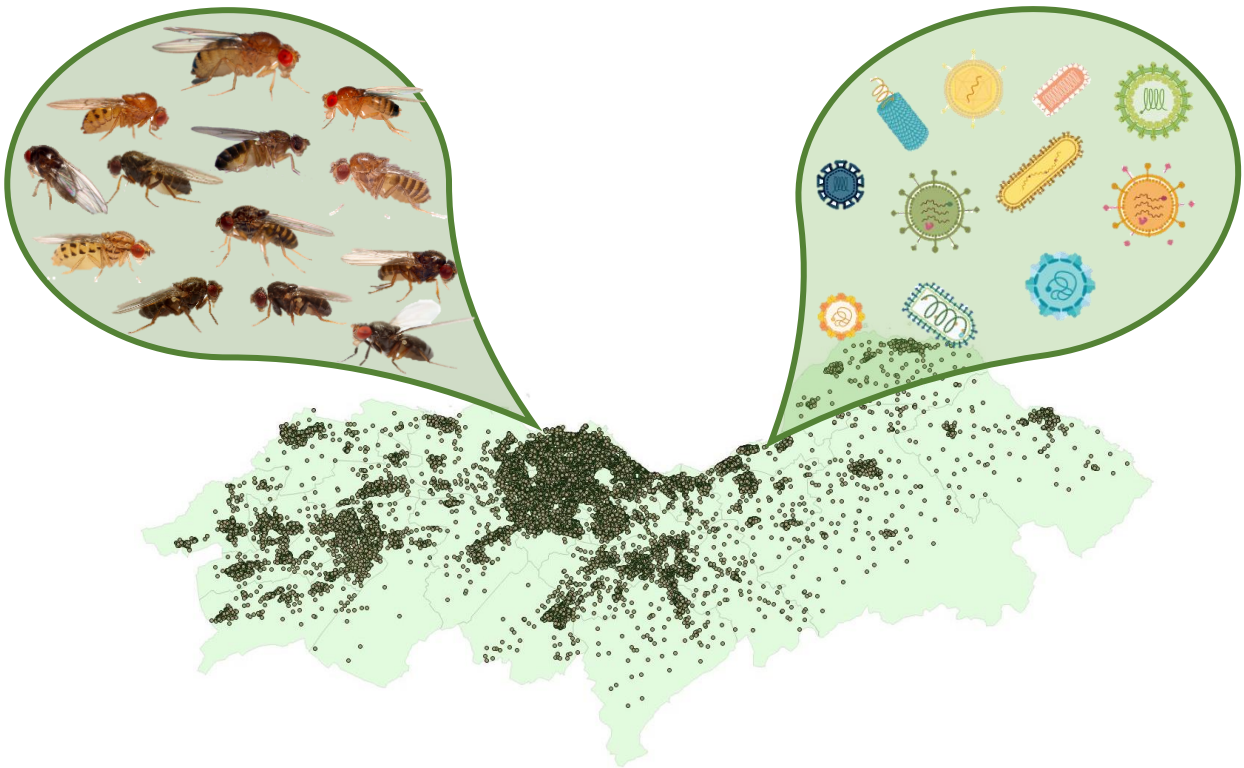
THE UNIVERSITY
of EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Virus discovery, dynamics, and disease in a wild *Drosophila* community

Megan A. Wallace



Doctor of Philosophy

Institute of Evolutionary Biology, University of Edinburgh

March 2021

Table of Contents

Abstract 8

Lay Summary 10

Acknowledgements 11

1 General Introduction: *Drosophila* and their naturally occurring viruses: a model system for insect-virus community dynamics and evolution 12

1.1. How many viruses infect insects? 15

1.1.1 RNA viruses 15

1.1.2 DNA viruses 16

1.1.3 Endogenous viral elements (EVEs) 17

1.1.4 Coinfection 18

1.2 How host specific are insect viruses, and are host switches common? 19

1.2.1 Host range 19

1.2.2 Host switching 21

1.3 How does insect virus prevalence vary in space and time? 22

1.4 How do insects and insect viruses evolve, or co-evolve? 24

1.5 How strong is insect virus-driven selection? 25

1.5.1 Infection phenotypes in *Drosophilidae* 27

1.5.2 Transmission routes of *Drosophila* viruses 28

1.6 Objective of the thesis 29

2 The discovery, host range and prevalence of RNA viruses infecting Scottish Drosophilids 31

2.1 Introduction 31

2.2 Methods 35

2.2.1 Collection and identification of multiple species of Drosophila 35

2.2.2 Time structured sequencing of mixed-species pools of Drosophilidae 36

2.2.3 Identifying virus-like contigs 37

2.2.4 Read mapping and viral genome organisation 38

2.2.5 *Characterising the host range of new and known Drosophila viruses* 40

2.2.6 Characterising variation in Drosophila virus prevalence and sequence across host species, space, and time 42

2.3 Results 45

2.3.1 Monthly collections of 15 Drosophila species over 3 years 45

2.3.2 Identification of known and new virus contigs, and their abundance across datasets 47

2.3.3 Host range of new and known *Drosophila* viruses 58

2.3.4 Prevalence of ten *Drosophila* viruses across species, season, and site 62
2.3.5 Genetic differentiation of *Drosophila* viruses across host species 65
2.3.6 Co-infections are not a rarity in wild multi-virus systems 66
2.4 Discussion 69
2.4.1 Virus discovery and read abundance in Scottish drosophilids 69
2.4.2 Drosophila virus host range 71
2.4.3 Variation in Drosophila virus prevalence by host species, and season 72
2.4.4 Ecological drivers of virus dynamics 73
2.5 Conclusions 75
3 Prevalence variation and genetic diversity in DNA viruses infecting European *Drosophila* 76
3.1 Introduction 76
3.2 Methods 81
3.2.1 Sample collection, sequencing and prior analyses 81
3.2.2 Spatial and temporal variation in viral copy number 82
3.2.3 Genetic diversity 84
3.2.4 Structural diversity in the Kallithea virus genome 86
3.3 Results 88
3.3.1 DNA virus prevalence varies in space and time 88
3.3.2 A recently integrated endogenous copy of Galbut virus is spatially and seasonally distributed 92
3.3.3 Genetic diversity 93
3.4 Discussion 98
3.4.1 Broad-scale spatiotemporal patterns of virus prevalence in DNA viruses of European *Drosophila* 98
3.4.2 An endogenous copy of Galbut virus likely had a Northern European origin and now shows seasonal patterns of incidence 101
3.4.3 Genetic diversity is variable, but population differentiation high in Drosophila-infecting DNA viruses 102
3.4.4 Patterns of evolution in DNA virus genomes 104
3.5 Conclusions 106
4 Viruses transmitted from wild *Drosophilidae* can reduce offspring number and lifespan in *Drosophila melanogaster* 107
4.1 Introduction 107
4.2 Methods 111
4.2.1 Exposure of Dmel to wild Drosophila 111
4.2.2 Detection of Drosophila virus infection & transmission events 113

4.2.3 Sanger sequencing of transmitted viruses 115
4.2.4 The impact of viral infection on Dmel lifespan 116
4.2.5 The impact of viral infection on Dmel offspring number 118
4.3 Results 122
4.3.1 Transmission of Drosophila infecting viruses to D. melanogaster 122
4.3.2 Four naturally-occurring *Drosophila* viruses significantly reduce lifespan in *D. melanogaster* 128
4.3.3 Three naturally-occurring *Drosophila* viruses significantly reduce lifetime offspring production from *D. melanogaster* females 134
4.4 Discussion 141
4.4.1 Cross-species transmission of Drosophila viruses 141
4.4.2 Infection with four viruses reduces lifespan in D.melanogaster 143
4.4.3 Infection with three viruses reduces lifetime offspring production in D.melanogaster 144
4.4.4 Caveats 146
4.5 Conclusions 148
5 General Discussion 149
5.1 Discovery of Drosophila associated viruses 149
5.2 The host range and prevalence of Drosophila viruses 150
5.3 The evolution of Drosophila RNA viruses in multi-host, multi-virus systems 152
5.4 Drosophila DNA virus prevalence, and diversity 153
5.5 The fitness costs of naturally-occurring Drosophila viruses 154
5.6 Conclusions 156
6 References 157
7 Appendix 174

Index of Figures & Tables

Figures

Main text

Fig. 2.1. Sampling strategy for Drosophila collections. 36

Fig. 2.2. *Drosophila* collection details. 46

Fig. 2.3. Genome organisation of newly described viruses. 54

Fig. 2.4. Phylogenies of newly described *Drosophila*-infecting RNA viruses.. 55

Fig. 2.5 Virus read numbers..... 57

Fig. 2.6. Host range matrix for newly described *Drosophila* RNA viruses. 60

Fig. 2.7. The host range Drosophila viruses. 61

Fig. 2.8 Drosophila virus prevalence across species..... 64

Fig. 2.9 The frequency of co-infection in wild *Drosophila*. 68

Fig. 3.1 Spatial variation in the copy number of Kallithea (A) and Linvill Road virus (B), and the Galbut EVE (C & D).. 90

Fig. 3.2. Comparison of the fixed effects estimates from each of the INLA models for viral prevalence.. 91

Fig. 3.3 The range of spatial autocorrelation acting on Kallithea virus, Linvill Road virus and the Galbut EVE. 92

Fig. 3.4. Neutral site-diversity (A) and support for short insertion deletion polymorphisms (B) across the Kallithea virus genome..... 97

Fig. 4.1. Summary of methods used to expose female Dmel to naturally-occurring Drosophila viruses.. ……..... 112

Fig. 4.2 DimmNV maximum clade credibility tree including experimental and wild collected flies.. 125

Fig. 4.3 Phylogenies of ImmSV sequences (regions of the L and N genes) from experimental and wild collected flies. 126

Fig. 4.4 The effects of viral infection on mortality over time in Dmel OreR females**.** 130

Fig. 4.5 Posterior density distributions of the viral fixed effects included in the model of lifespan variation in Dmel females. 132

Fig. 4.6 The effect of viral infection on lifespan in female Dmel (OreR). 133

Fig. 4.7 The effect of the number of viral infections on lifespan in female Dmel.. 134

Fig. 4.8 No. of offspring produced by Dmel females, over time, when infected with viruses. 136

Fig. 4.9 Posterior density distributions of the viral fixed effects included in the hurdle-Poisson model of lifetime offspring production in Dmel females**.** 139

Fig. 4.10 The impact of viral infection on lifetime offspring production in female Dmel (OreR). 140

Appendix

Fig. S2. 1 Read mapping against genomes of newly described *Drosophila* viruses. 185

Fig. S2. 2 Genome organisation of extended *Drosophila* viruses.. 187

Fig. S2. 3 Host range matrix by RT-PCR for newly described *Drosophila* RNA viruses, including all segments. 188

Fig. S2. 4. Phylogenies of multi-host viruses..... 190

Fig. S2. 5. Prevalence of *Drosophila* viruses by month 191

Fig. S2. 6. Prevalence of *Drosophila* viruses by site.. 193

Fig. S4. 1 Potential issues with sampling lifespan with RNA virus clearance…… 207

Fig. S4. 2. Predicted number of Dmel females producing zero offspring under the chosen hurdle model for lifetime offspring production. 208

Fig. S4. 3 Posterior density distributions of the viral fixed effects included in the model of lifespan variation in Dmel females, with extra data for DmelNV. 209

Fig. S4. 4 Posterior density distributions of the viral fixed effects included in the model of lifetime offspring production, and early life offspring production in Dmel females. 210

Fig. S4. 5 Simulation of power to detect a viral-induced reduction in lifespan. 211

Tables

Main text

Table 2.1. New viruses reported in this chapter.. 53

Table 2.2. Log likelihood of viral prevalence variation under four models..... 63

Table 3.1 Genetic diversity of three DNA viruses infecting European Drosophila..... 94

Table 4.1 Summary of viral exposure and transmission from wild flies to OreR Dmel. 121

Table 4.2 Outputs from linear mixed effects models used to analyse the effect of viral infection on lifespan in female Dmel..... 131

Table 4.3: Outputs from the hurdle-Poisson mixed effects model used to analyse the effect of viral infection on lifetime offspring production in female Dmel. 138

Appendix

Table S2 1 Sampling sites used to collect *Drosophila*…….. 175

Table S2 2. RT-PCR Primer assays designed for newly described RNA viruses. 178

Table S2 3 Known Drosophila virus PCR primers..... 182

Table S2 4. Extended *Drosophila* virus genome assemblies**.**..... 184

Table S3. 1 Total population πA, πS and πA/πS for each gene in the genomes of Kallithea, Linvill Road and Vesanto virus (including all haplotypes)……. 199

Table S4. 1 PCR primers for scanning wild-collected flies…… 204

Table S4. 2 PCR primers for scanning female Dmel……. 206

Abstract

The fruit fly *Drosophila melanogaster* remains a key model system for the study of insect-virus interactions. Its tractable life history and associated genetic toolkit have aided in the discovery of many of the key invertebrate immune defences against viruses. Recent metagenomic sequencing studies have identified over 130 naturally-occurring viruses of the *Drosophilidae*. This could enable the study of insect-virus ecological and evolutionary dynamics in this native multi-host, multi-virus community. In particular, the use of naturally-occurring host-virus combinations allows the study of 'typical' wild co-evolutionary dynamics between insects and their viruses, which might help us to understand the evolution of insect-vectored viruses with economic and public health impacts. However, proper parameterisation of these models of insect-virus co-evolution requires data from long-term studies of wild host-virus communities, which do not currently exist.

In this thesis, I make the first attempt to quantify the dynamics of wild *Drosophilid* virus communities at both local, and continent-wide scales. I begin by using metagenomic sequencing to describe the viruses present in populations of *Drosophila* in South-east Scotland, identifying 17 new RNA viruses. I characterise the host range of 41 new and previously described *Drosophila* viruses in this system, finding that over 90% of these viruses infect multiple host species. I then examine how ten *Drosophila* viruses vary in prevalence in the wild using repeated, spatially and temporally structured sampling of a total of 2227 flies of 15 species, over three years. I find that prevalence varies between viruses, and within viruses, can vary across host species, and collecting season. Co-infection is not a rarity in this system, as over 30% of flies are infected with multiple viruses. Expanding my spatial scale of analysis, I then characterise patterns of *Drosophila melanogaster* DNA virus prevalence and genetic diversity across Europe. I find that DNA viruses show varying levels of diversity, and large-scale spatiotemporal predictability. Finally, I experimentally examine the ability of *Drosophila* viruses to transmit across species, and quantify their impact on host fitness. I find that some

Drosophila viruses can readily transmit across species without the need for systemic injection. I identify four *Drosophila* viruses which significantly reduce host lifespan, and three which significantly reduce offspring production, indicating that viral infection may be a substantial fitness burden on wild flies. Together, these data increase and demonstrate the utility of the *Drosophila* model for community-level studies of host-virus interactions.

Lay Summary

Some viruses that cause human diseases, like dengue virus and Zika virus, are transmitted by insects. Insects can also carry viruses that they don't share with humans or other animals, but which still harm insects with economic or ecological value. One example of this is Deformed wing virus, which is currently causing declines in honey bee populations around the world. To understand why insect viruses like this one emerge and devastate insect populations, we can study 'model' insects and their viruses in the wild.

In the last decade, we've gone from knowing very little about what viruses infect insects, to knowing that the insect-infecting viruses are as diverse as those which infect other animals and plants combined. One particularly well-studied group are fruit flies, which carry over 100 different viruses in the UK. The vinegar fly, *Drosophila melanogaster*, has been used in laboratory experiments to find out how insects defend themselves against viruses. We could combine this knowledge of their defence mechanisms with our knowledge of their wild viruses, to ask why, in wild systems where insect species live alongside one another, some viruses are shared between them, and some are not.

In this thesis, I have explored which viruses infect fruit flies in the U.K, and why some viruses can only infect one insect species, while others infect many. I have also looked at how the presence or absence of fruit fly viruses varies across the wider European continent, why fruit flies in some European regions have lots of viral infections, and how diverse these viruses are. I finish by looking at what these viruses actually do to the fruit flies they infect. I ask whether they can reduce the lifespan, or the number of offspring produced by flies. If they do, this means that these viruses could drive natural selection in insect populations.

Acknowledgements

I'd like to first thank my supervisor, Dr. Darren Obbard, for his patient and attentive supervision of me the past four years. You pushed me to become a better scientist and critical thinker, and I hope I've at least in part succeeded. Thanks also go to the other members of the Obbard lab, for keeping me company in the lab during many, many PCRs, and for not complaining about my choice of podcasts. I'd also like to thank my thesis committee, Dr. Pedro Vale and Prof. Tom Little for their helpful advice and discussions in my annual review meetings. Thanks also to the rest of the Regan, Walling and Vale labs for so many useful discussions about my work at our weekly meetings, and excellent company at the pub.

This thesis wouldn't have been possible without a whole host of people and organisations who gave me permission to come and collect flies on their land on a monthly basis, thank you to you all.

To my friends, who I have missed so much this year, thank you for all your support, advice, and general steerage of my over-thinking mind. Particularly thanks to Katie, who made me many cups of tea and significantly improved the quality of this thesis by being such a supportive flatmate. TYD, you've made my Ph.D. years 200% more enjoyable. I couldn't wish for a more inspiring, supportive and hilarious group of friends. If we're all friends until we're old and can afford to buy a commune with a hot tub, I'd be ok with that.

To my family, my apologies for being a terrible daughter/sister/granddaughter/niece the past year, and for all the times I didn't reply to your text or email. Truly, I can't thank you all enough for your support. In particular, thank you to my Mum and Dad, who took me in for 7 months when my U.S placement plans fell through in a global pandemic, let me borrow the car for fieldwork on a monthly basis, and even read one of my papers! I promise this is my last degree.

1 General Introduction: *Drosophila* and their naturally occurring viruses: a model system for insect-virus community dynamics and evolution

Insect viruses and their hosts' immune systems are assumed to be engaged in reciprocal, co-evolutionary arms-race dynamics. This assumption is supported by the fast evolution of some insect immune genes (Obbard *et al.* 2006), and the evolution of large-effect segregating polymorphisms for resistance (Magwire *et al.* 2012; Cao *et al.* 2016). However, most *in situ* studies of reciprocal host-virus co-evolution are on single-host, single-pathogen systems, where it is assumed that the evolution of host immune genes can compete with the evolution of immune evasion by the virus. In reality, most insect-virus co-evolutionary interactions occur in complex, multi-host, multi-virus communities (Hall *et al.* 2020), where ecological dynamics also influence the ability of host and virus to exert selection on the others genome.

Multi-host systems are the norm in nature, so when we examine single-host, single-virus interactions, we may simply be examining a tiny portion of the diversity of virus-driven selection imposed on the host immune system, and a tiny portion of the diversity of host driven selection imposed on the virus. To rectify this, and conduct subsequent studies of host-virus coevolution, we would have to either measure, or accurately predict, the number of viruses infecting each insect host species, and the host range of each virus. In reality, we will need a combination of both approaches. Firstly, we will need an approach that utilises model systems, where we can intensively sample virus host range and prevalence across species, while controlling for sampling effort, and environmental and sampling derived auto-correlation. Then, with this data from model systems, we could parameterise models which predict virus host range, switching, and evolution across the insect phylogeny. Here, a familiar model system presents itself as a candidate for such studies, *Drosophila*. The fruit fly *Drosophila melanogaster* (Dmel), along with its close relatives, presents a powerful model system in which to investigate these viral dynamics at both a global and a local scale.

The immune system of Dmel (Diptera, family *Drosophilidae*), is one of the best characterized of the metazoans, and has been well studied as a model of the innate immune response in both invertebrates, and vertebrates (Buchon *et al.* 2014; Neyen *et al.* 2014). For example, the identification of the Toll protein of *Drosophila* as an important defence against bacterial infections led to the identification of Toll Like Receptors in vertebrates, as a key component of their innate immunity (Takeda & Akira 2005). In the search for mechanisms underlying variation in viral interactions with their hosts, Dmel has also provided insights into mechanisms of insect anti-viral defence.

The three key components of *Drosophila* anti-viral defence are RNA interference (RNAi), inducible responses, and cellular responses. The first component, the RNAi response, which is conserved across plants (Ratcliff *et al.* 1997) and other animal phyla (Fire *et al.* 1998), acts to inhibit the expression of viral proteins through small interfering RNAs (siRNAs) (reviewed in Mussabekova *et al.* 2017). The importance the RNAi response in insect anti-viral defence is demonstrated both in viruses, and in flies. In viruses, by their expression of RNAi suppressors (RNA virus - Nayak *et al.* 2010; DNA virus - Bronkhorst *et al.* 2014) and in flies, by their increased sensitivity to RNA and DNA viruses on the loss of function of key RNAi genes (Wang *et al.* 2006), and the accumulation of virus-derived siRNAs on infection (Wang *et al.* 2006; and use of this to discover viruses in Webster *et al.* 2015). The second key component of Drosophila anti-viral defence is inducible responses. Inducible anti-viral responses consist of three evolutionarily conserved inflammatory pathways; two signal transduction pathways: the nuclear factor κB (NF-κB) pathways (Toll and IMD), and one cytokine-activated pathway: the Jak/STAT pathway. Toll pathway genes are required for resistance to oral infection by several RNA viruses (Ferreira *et al.* 2014), and the importance of the IMD pathway to anti-viral defence is shown by the acquisition of a negative regulator of the IMD pathway by several DNA viruses of insects (Lamiable *et al.* 2016). The Jak/STAT pathway appears to contribute to virus specific defences against *Dicistroviridae* (Kemp *et al.* 2013). The third key component, cellular processes, such as phagocytosis, apoptosis, and

autophagy can also limit viral replication, dissemination and eventually transmission in insects. The power of apoptosis to limit viral replication per cell is demonstrated by the acquisition of a cytokine which antagonises apoptosis by DNA viruses of insects (Mlih *et al.* 2018). Additionally, polymorphisms in the *ref*(2)P gene, part of the autophagy pathway, are significantly associated with resistance to Drosophila melanogaster sigma virus infection (Magwire *et al.* 2012). By combining our knowledge of the *Drosophila* anti-viral immune responses, with studies of the ecological dynamics of insect-virus systems, we could study the context in which these anti-viral responses arose in wild populations.

The utility of the sympatric *Drosophila* species in characterising insect-virus interactions has also been demonstrated by studies such as Van Mierlo *et al.* (2014). In this study, the authors characterised the host-specific nature of immune response suppression by three naturally-occurring *Drosophila* viruses across multiple *Drosophila* species. They demonstrated not only the value of having working in vitro assays for your host species, but the flexibility this gives you to ask questions about the molecular mechanisms underlying complex ecological patterns such as host range. These assays and techniques can be combined with our expanding knowledge of the genomic diversity present in the wider genus *Drosophila*. 101 new *Drosophilid* genomes were recently assembled from both short and long read sequencing (Kim *et al.* 2020), and a dataset of 155 *Drosophilid* genomes compiled (Suvorov *et al.* 2021) which could be used for comparative studies of genetic variation, and selection in insect anti-viral immune genes.

Using the *Drosophila* model system, we can ask key questions about the context of insect-virus interactions, such as;

1) How many viruses infect insects?

2) How host-specific are they, and are host switches common?

3) How does insect virus prevalence vary in space and time?

4) How do insects and insect viruses evolve, or co-evolve?

5) How strong is virus-driven selection?

In this chapter, I will present our current understanding of the answers to these questions, the insights already gained from the *Drosophila* model, and highlight areas of research where we can further utilise this system to model multi-host multi-virus community dynamics.

1.1. How many viruses infect insects?

1.1.1 RNA viruses

Historically, insect viruses were described which caused disease, or from species with economic impacts. The earliest arthropod-borne RNA viruses described, when virus discovery was reliant on cell culture isolation, were two flaviviruses, yellow fever virus in mosquitoes, and Louping ill virus in ticks (Calisher & Gould 2003). However, in the past two decades, lowering sequencing costs have enabled metagenomic studies of a wider range of the arthropod phylogeny, filling in gaps in our knowledge of not only arthropod viruses, but RNA virus diversity in general (see review Greninger 2018). In insects, and the wider invertebrates, large scale transcriptomic studies have exponentially expanded our knowledge of the number and diversity of viruses infecting these groups. For example, Li *et al.* (2015) presented the transcriptomic profiles of 70 arthropod species, identifying 112 new negative sense RNA viruses, and describing a new virus family, the *Chuviridae*. Importantly, they found that much of the diversity of plant, and vertebrate negative sense RNA viruses lies within the diversity of arthropod associated viruses. It's possible that insect's large, sometimes dense, populations foster virus diversity, and that their close interactions with plants and vertebrates spread this diversity across other phyla. The same research group then used transcriptome sequencing of over 220 invertebrate species, and nine animal phyla, to describe 1445 new RNA viruses (Shi *et al.* 2016a). This took their exploration of virus diversity outside of the negative sense RNA virus genomes, adding not only two new negative sense RNA virus families to the overall virus phylogeny, but also three positive sense RNA virus families. Since these landmark studies, many others have continued to expand our knowledge of the insect virosphere (Zhang

et al. 2019b; Obbard *et al.* 2020), and now, RNA virus discovery is becoming even more accessible to researchers working on insects, with the release of common protocols (Viljakainen & Jurvansuu 2020).

Concurrent with all of this rampant virus discovery in the wider invertebrates, the virome of the *Drosophilidae* also benefited from metagenomic, and meta-transcriptomic analyses. Prior to the advent of metagenomic sequencing, the viruses used in *Drosophila* experiments had been largely discovered as laboratory contaminants (eg. Plus & Duthoit 1969), or through distinctive laboratory phenotypes (eg. sigma viruses on CO_2 - L'Heritier & Teissier 1937). However, the late 2000s saw an effort to expand our knowledge of the diversity of naturally occurring RNA viruses infecting this model organism, and its close relatives. This first was focussed on the wisely used Sigmavirus genus, with Longdon *et al.* (2010) identifying and sequencing at least the polymerase gene of two new viruses: Drosophila obscura sigmavirus and Drosophila affinis sigmavirus. Subsequently, Webster *et al.* (2015) used a metagenomic survey of ~2000 wild Dmel and *Drosophila simulans* (Dsim) collected in Europe, Africa, North America and Australia to identify over 20 new RNA viruses associated with this species. They also used the presence of virus-derived siRNAs to confirm that these were active infections of *Drosophila.* The *Drosophila* virosphere was then expanded upon in a metagenomic study of six *Drosophila* species native to the UK that identified 25 more Drosophilid-infecting viruses (Webster *et al.* 2016). These viruses included close relatives of some of the non-native *Drosophila* viruses already used in experiments, such as Newington virus, which is closely related to Flock house virus (Dasgupta & Sgro 1989), originally from a coleopteran. Since then, further studies have identified 18 new RNA viruses that infect the soft-fruit pest *Drosophila suzukii* (Medd *et al.* 2018a), and close relatives of *Drosophila* associated RNA viruses have been identified infecting other insects (eg. Cross *et al.* 2020; Sharpe *et al.* 2021).

1.1.2 DNA viruses

The host range, ecology and evolution of double stranded DNA (dsDNA) baculoviruses have been extensively studied in Lepidoptera and other insect species (reviewed in Cory & Myers 2003). However, prior to the era of metagenomic sequencing we had little information on the DNA viruses infecting most other invertebrate groups, though some studies had documented the abundance of DNA bacteriophage in the early 2000s (Rohwer 2003). Since then, the use of metagenomic methods has led to the description of DNA viruses across a wide range of invertebrate orders, such as dsDNA Hytrosaviruses in Diptera (Kariithi *et al.* 2017) and Rep-encoding single stranded DNA (ssDNA) viruses (CRESS DNA viruses) in all three major terrestrial arthropod lineages (Rosario *et al.* 2018). These ssDNA viruses with small circular genomes have been primarily described through shotgun sequencing of purified viral particles, and rolling circle amplification, methods which bias discovery towards these types of viruses (Kim & Bae 2011; Delwart & Li 2012). Meta-transcriptomic sequencing of ~220 species across nine invertebrate phyla has also recently described 13 new ssDNA viruses, and 6 dsDNA viruses (Porter *et al.* 2019).

Metagenomic sequencing studies in *Drosophila* have so far detected far more RNA viruses than DNA viruses, despite RNA sequencing being able to 'discover' DNA viruses, if they are being actively transcribed. This suggests that, in *Drosophila*, DNA viruses are far rarer than RNA viruses. Prior to 2021, seven naturally-occurring DNA viruses of *Drosophila* had been described. These comprised three large dsDNA Nudiviruses, a large dsDNA Iridovirus, two ssDNA Densoviruses, and a putative ssDNA Bidnavirus (Unkless 2011; Webster *et al.* 2015, 2016; Kapun *et al.* 2020). Some of these Drosophila associated DNA viruses encode suppressors of anti-viral responses (eg. Bronkhorst *et al.* 2014), suggesting a history of co-evolution with their hosts. For example, Kallithea virus encodes a potent suppressor of Toll signalling through the regulation of NF-κB transcription factors, a mechanism which is also active during *Drosophila innubila* Nudivirus infection, and therefore appears to be evolutionarily conserved (Palmer *et al.* 2018a).

1.1.3 Endogenous viral elements (EVEs)

Metagenomic sequencing studies have not only identified a large and diverse virosphere associated with insects, but also DNA sequences from these viruses integrated into insect genomes. The integration of viral sequences into the host genome can occur in both somatic and germline cells, and can originate from DNA viruses (eg. Kimenyi *et al.* 2020), retroviruses (eg. Holt *et al.* 2002), and non-retroviral RNA viruses (eg. Ballinger *et al.* 2014). *Drosophila* genomes contain examples of these integrated non-retroviral RNA viruses (Ballinger *et al.* 2012). These EVEs might even play some role in protecting the host against viral challenge. In mosquitoes, non-retroviral EVEs have been observed to often associate with transposable elements and occur in PIWI-interacting RNA (piRNA) generating clusters (Palatini *et al.* 2017a). Indeed, piRNA clusters seem to be associated with integrated ssRNA virus genomes across arthropods (ter Horst *et al.* 2019), and piRNAs generated from these endogenous viral sequences have even been suggested to provide some immunity to exogenous viruses (eg. Tassetto *et al.* 2019; Joosten *et al.* 2020).

1.1.4 Coinfection

By doing a detailed assessment of the number of viruses that infect specific insects we could also gain insights into the rates of viral co-infection in wild insect populations. Co-infections may be the norm, as a recent metatranscriptomic study of single mosquitoes found that ~88% were infected with multiple viral taxa (Batson *et al.* 2020). However, there are currently few studies of the impact that co-infection can have on insect host fitness, or susceptibility to viral infection (reviewed in Vogels *et al.* 2019 for arboviruses). When this review calculated the relative transmission rates of singly-infected and co-infected mosquitoes with arboviruses across studies, they found differences in transmission between single and multiple infections to be generally small. In contrast to this, there are more studies on the impact of viral co-infection in plants, which, in particular, highlight the importance of the sequential nature of co-infections for their outcome in wild hosts (eg. Marchetto & Power 2017), a point reiterated by studies of co-infecting Microsporidia in honeybees (Natsopoulou *et al.* 2015).

Theoretically, viral co-infection could result in three outcomes: viral antagonism, viral facilitation or viral accommodation. Studies in mosquitoes suggest that if co-infecting viruses are competing for the same host resources, they can inhibit the growth of one another within cell culture (Schultz *et al.* 2018) or individuals (Vazeille *et al.* 2016), leading to competitive displacement by, or non-establishment of the second virus. However, when pre-existing viruses are able suppress the host immune response (eg. viral suppressors of RNAi) they might amplify the virulence of a second viral infection: viral facilitation. There are few examples of this kind of interaction, but co-infection of a mosquito cell line with Japanese encephalitis virus and *Culex* flavivirus caused severe cytopathic effects (Kuwata *et al.* 2015). We know that in some cases, mosquitoes *are* able to mount a specific RNAi response to three, simultaneous, co-infecting viruses (eg. Frangeul *et al.* 2020), but this might depend on none of the viruses inhibiting the host immune response. Alternatively viral co-infection might have no effect on host fitness, or susceptibility to viruses at all. This lack of an effect on host fitness would be tricky to characterise with certainty, but co-infection of mosquitoes with Zika and Chikungunya virus does not affect vector competence, demonstrating viral accommodation due to the different subcellular niches occupied by the two viruses (Göertz *et al.* 2017). The effects of interacting viral infections have not, so far, been tested in the *Drosophila* model, despite >20% of wild-caught flies carrying multiple RNA viruses (Webster *et al.* 2015; Shi *et al.* 2018). With the variety of naturally occurring viruses of *Drosophila* now available for this research, with different replication systems, levels of immune suppression and virulence, this system provides a great resource for examining what influences the outcome of viral co-infection in insects, and how frequent co-infections are in the wild.

1.2 How host specific are insect viruses, and are host switches common?

1.2.1 Host range

Most viruses are capable of infecting multiple host species. Some viruses constitute true multi-host pathogens, infecting and transmitting between multiple species, in contrast to others,

where rare ‘spillover’ events into a second host species end in dead-end infections (Fenton & Pedersen 2005). However, the identity of these host species, both reservoir hosts and spillover recipients – viral host range - is a difficult trait to characterise. This is because a complete characterisation of the host range of most viruses requires the exhaustive sampling of many host species to confirm presence/absence. Even for the best studied viruses, such as Influenza viruses and Ebola virus, host range likely remains incompletely characterised (Caron *et al.* 2018; Long *et al.* 2019). This matters because the determinants of host range underlie the patterns of incidence and transmission seen for zoonotic diseases with large economic and public health impacts (Woolhouse & Gowtage-Sequeria 2005). Characterising the host range of insect viruses can also help us to determine what restricts some viruses (insect specific viruses) to their biting insect hosts (eg. Junglen *et al.* 2017) while others can infect vertebrates, and plants. For those viruses which do transmit outside of insects, quantifying the relative contribution of different insect species to virus transmission and maintenance in multi-host communities (eg. in gatrointensinal parasites - Fenton *et al.* 2015) could be key to disease control.

In well-studied insect species, such as honey bees, some RNA viruses, such as deformed wing virus, are shared across diverse host species (Levitt *et al.* 2013), have a wide geographic range (Galbraith *et al.* 2018), and can be shared between managed colonies and wild bee populations (Tehel *et al.* 2016). However, wild insect viruses can also display high rates of specialism to specific hosts when sampled in smaller geographic regions. For example, a survey of mosquitoes in a range of habitat types in Côte d’Ivoire found that their associated viruses showed high host-specificity, with 39/49 viruses identified infecting only one of 42 mosquito species (Hermanns *et al.* 2021). The sustainability of this viral strategy is likely dependent on the consistent presence of its host species.

Many viruses of *Drosophila* infect multiple host species. In the laboratory, the phylogenetic history of some viruses suggests frequent cross-species infections. For example, *Drosophila* C virus was identified infecting eight *Drosophila* species with a common ancestor dating back

63 million years (Kapun *et al.* 2010). Evidence from wild *Drosophila* communities suggests that host generalism could also persist there. *Drosophila innubila* Nudivirus (DiNV) is both taxonomically and geographically widespread, infecting *Drosophila* species from 3 subgenera, from both the north-eastern and south-western United States (Unkless 2011). Additionally, sampling of British Drosophilids in 2015 and 2016 found that 61% of RNA viruses assayed for infected multiple host species (Medd 2019, unpublished thesis data), The reasons why the viruses of *Drosophila*, and other insects, display varying levels of host specificity remains unresolved.

1.2.2 Host switching

The initiation of a new long term host-virus association usually occurs through a host-shift. This is when a virus infects a non-reservoir species: a species that it has not been maintained in and transmitted among the population of long term (eg. Wallace *et al.* 2014). These cross species transmission events are ubiquitous feature of the phylogenetic history of, in particular, RNA viruses (Geoghegan *et al.* 2017). When they lead to sustained transmission within the new host species population, a successful host shift has occurred. However, a large proportion of these events fail to cause onward intra-species transmission, constituting epidemic fade-out or a dead end infection, and are known as spillover (Parrish *et al.* 2008; Longdon *et al.* 2014). Ultimately, the complex interactions of ecological and phylogenetic factors that drive the frequency of host shifting remain incompletely resolved, despite their sometimes extensive burden on public health (eg. Keele *et al.* 2006; Dos Reis *et al.* 2009; Andersen *et al.* 2020).

For RNA viruses, which show a particularly high propensity for shifting between distantly related hosts, multiple studies in mammals have identified host geographic range overlap, and host species relatedness as significant predictors of viral sharing (Davies & Pedersen 2008), and the probability of a virus successfully establishing itself in a new host species (Streicker *et al.* 2010; Faria *et al.* 2013). Across mammalian species, large datasets have enabled phylogeographic predictive studies of the patterns of viral sharing, finding that, for vector-borne

RNA viruses in particular, host ‘space sharing’ plays a key role in driving observed patterns (Albery *et al.* 2020). In the future, more machine learning based predictive models (eg. Babayan *et al.* 2018) could also be used to identify the arthropod vectors of ‘orphan’ viruses, but these models require training on known host-virus combinations, which are less well characterised for insects.

In the wild, the phylogeny of *Drosophila* sigmaviruses shows evidence of host shifting (Longdon *et al.* 2011c), concurrent with the frequent host shifting observed in the phylogeny of tephritid fruit fly viruses (Sharpe *et al.* 2021). In the lab, the *Drosophila* model system has also helped to elucidate the determinants, and consequences of host switching by their viruses (Longdon *et al.* 2011a, 2015a). An experiment investigating the specificity of three Sigma viruses to their natural hosts across 51 *Drosophila* species found that closer relatives of the virus’ natural host support higher viral replication rates after a host shift (Longdon *et al.* 2011a). It is likely that the vertical transmission route of sigma viruses (family *Rhabdoviridae*), which may be rare amongst the viruses of the *Drosophila,* contributes to this host specificity. But there is still a strong phylogenetic component to the distribution of viral load when a faecal-orally spread RNA virus, *Drosophila* C virus, is used in cross-infection experiments (Longdon *et al.* 2015a). Host phylogeny also influences the evolution of this virus in novel hosts, with mutations that adapt the virus to its new host also predisposing it to infecting its new hosts’ close relatives (Longdon *et al.* 2018). A further experiment, using four isolates of this same genus, *Cripavirus*, found tentative evidence that more closely related viruses are more correlated in their ability to infect a novel host species (Imrie *et al.* 2021), but a larger number of viruses would be needed to test this hypothesis conclusively. Fundamentally, the reasons why even some closely-related insect host species differ in the diversity, and abundance of viruses they carry (eg. Shi *et al.* 2017) remains unresolved.

1.3 How does insect virus prevalence vary in space and time?

In vertebrate-associated viral infections, virus incidence, and prevalence varies significantly in time and space (eg. O'Brien & Xagoraraki 2019), and research in single, well-studied communities, such as *Drosophila*, honey bees and mosquitoes (eg. Fei *et al.* 2011; Webster *et al.* 2016; Batson *et al.* 2020), suggests that the same is true for insect viruses. Seasonal fluctuations in host density are predicted to drive an increase in prevalence of these viral diseases in temperate climates (Anderson & May 1980), and in tropical climates, host density can still drive differences in virus prevalence. For example, host density affects the prevalence of salivary gland Hytrosaviruses (SGHVs), but interestingly, the direction of this effect varies with transmission route. When host density increases, vertically transmitted SGH in tsetse flies shows lower prevalence (Odindo 1982), but orally transmitted SGHVs in house flies shows higher prevalence (Geden *et al.* 2008). Viruses infecting the same insect species can also show marked differences in their seasonal prevalence distributions, demonstrated by the monthly sampling of six viruses, infecting two species of honey bee across the U.S (Runckel *et al.* 2011). This suggests that the burden of viral infection on a wild insect population does indeed vary significantly in time, and that the pattern of this variation is specific to the hosted viral community.

Insect virus prevalence can also vary in space. Geoghegan *et al.* (2014) analysed the seasonality of three arboviruses of cattle using a sero-surveillance dataset collected over 8 years. They found that both the timing of epidemics and the average number of seroconversions has a strong geographic component. Anthropogenic habitat changes could also drive changes in insect virus prevalence, as Myer & Johnston (2019) found that the incidence of West Nile virus in mosquito traps not only showed significant spatial variation, but also was higher in areas of low vegetation and urban development. Habitat disturbance, which often drives an increase in the abundance of resilient species, can also drive changes in the community composition and abundance of their viruses (Hermanns *et al.* 2021). Many of these seasonal, and spatial trends are also likely to be driven by environmental factors. For example, the prevalence of the SGH viruses is positively correlated with vegetation density and rainfall,

and negatively correlated with temperature (Odindo & Amutalla 1986). However, separating these environmental factors from spatiotemporal trends, to test their power to drive changes in virus prevalence is statistically challenging, and requires repeated, spatiotemporally structured sampling of insect-virus communities.

The prevalence of *Drosophila* viruses varies significantly across a global scale, as evidenced by virus presence/absence observations from 17 global locations in (Webster *et al.* 2015), and the varying prevalence of *Drosophila* innubila Nudivirus in quinaria group species from the Chiricahua Mountains (~60%), and lower prevalence (~3%) in *D. falleni* (Unkless 2011). *Drosophila* virus prevalence is also expected to vary significantly over a smaller seasonal and temporal scale, coupled with the seasonal variation of their associated host species (Bombin & Reed 2016). However, no studies have yet examined local-scale variation in *Drosophila* virus prevalence, or the significance of environmental factors in driving this variation.

1.4 How do insects and insect viruses evolve, or co-evolve?

Co-evolution - reciprocal, adaptive genetic change in two or more interacting species (Woolhouse *et al.* 2002) - could be an integral driver of genetic change in insects and their viruses. Because of the intimate, and (for the virus) dependent nature of the association, this selection pressure is predicted to be particularly strong in pathogens and their hosts. A pathogen which decreases host fitness should drive selection for resistance, and, in the virus, reciprocal selection for repressors of this immune response, creating genotype by genotype interactions. However, evidence of these strict, reciprocal, co-evolutionary interactions are difficult to find in invertebrates.

Insect-driven selection on viruses is likely to underlie the host specificity seen in some insect viruses. For example, *Anopheles* specific flaviviruses display a nucleotide motif bias in their genomes consistent with evolution in insect hosts (Colmant *et al.* 2017b), a pattern also present in the genomes of other insect specific flaviviruses (Lobo *et al.* 2009). Viral

suppressors of the host immune response that interact with protein components of the RNAi response can also be extremely host-specific (van Mierlo *et al.* 2014). Some viruses also show evidence of specific adaptive changes, which allow them to better exploit their hosts. Adaptation to infect an increased range of host species, or to achieve higher viral titre has been attributed to specific amino acid changes in several invertebrate vectored viruses; such as Chikungunya virus (Tsetsarkin & Weaver 2011), and a virus of *Drosophila,* Drosophila innubila Nudivirus (Hill & Unckless 2020). It's possible that other genes which display signatures of positive selection in insect genomes also are targets of host suppression, such as genes of the *Baculoviridae*, dsDNA viruses which infect a diverse range of insect species, involved in replication, and which display signatures of recurrent positive selection (Hill & Unckless 2017).

In insect genomes, virus-driven selective pressure could manifest as positive selection on anti-viral genes, negative frequency dependant selection, or selective sweeps on large effect polymorphisms. Rapid adaptive evolution has been observed on some insect anti-viral genes, such as the RNAi genes of *Drosophila* (Obbard *et al.* 2006; Kolaczkowski *et al.* 2011). Large effect polymorphisms for host resistance have also been identified in *Drosophila* in the lab (Magwire *et al.* 2012), and in the wild (Wayne *et al.* 1996), providing evidence of virus-driven selection in insects. However, there is now a recognition that coevolution in nature occurs within complex ecological networks of multiple species, and it remains challenging to separate directional and fluctuating selection dynamics, which would purge or maintain genetic diversity respectively, in these systems (Hall *et al.* 2020). Local adaptation in multi-host communities may increase the specificity of pathogen genotypes for their hosts, as seen in a plant-pathogen community (Chappell & Rausher 2016). However, the specialism of specific viral genotypes to their hosts has not been examined at a local level for wild multi-host communities of insects.

1.5 How strong is insect virus-driven selection?

These kinds of active co-evolutionary dynamics are only predicted to occur in response to a virus that reduces the evolutionary fitness of its host. This is usually measured as a reduction in lifespan, but this only matters in the context of selection if it reduces the offspring produced by the infected insect. To evaluate the strength of virus-driven selection on insect immune systems, we need to look at the reduction in evolutionary fitness that insect viruses cause. Like metagenomic sequencing efforts, studies characterising the pathology of insect viruses are far more common in crop-pests, plant disease causing viruses vectored by insects, insects with economic value such as honeybees, and vectors of human-infecting viruses, such as mosquitoes. Taking mosquitoes and *Drosophila* as examples, we can examine the current knowledge and difficulties in characterising these viral effects on insect fitness.

Infection with several viruses has been found to reduce the lifespan of mosquitoes, including West Nile virus, dengue virus, chikungunya virus and Eastern equine encephalitis virus. More recently, work by Sedger *et al.* (2018) found that infection with Zika virus can reduce the lifespan of its primary vector, *Aedes aegypti,* but does not detectably affect female ovipositional success or eggs per clutch. Fewer viral infections have been directly linked with reductions in fecundity, but West Nile virus has also been found to reduce the fecundity of colonised *Culex tarsalis* mosquitoes (Styer *et al.* 2007). However, often mosquito vector populations and virus strains display a high degree of species-specific differences in their pathology, transmission dynamics, and costs of resistance. The importance of using naturally-occurring, and local strains of a virus in experiments is demonstrated by the reduced Zika virus vector competence seen when Chouin-Carneiro *et al.* (2016) infected Brazilian mosquitoes with an Asian virus genotype, in comparison to local strains of the virus. Another example is found in the species specific effects of West Nile virus on *Culex* species fecundity, reducing *Culex tarsalis* fecundity (Styer *et al.* 2007) but showing no effects on fecundity in *Culex pipiens* (Ciota *et al.* 2011). So, in mosquitoes, viruses can impose tangible fitness costs on their hosts but these costs may vary in a species-specific manner, highlighting the need for the use of naturally-occurring host-virus combinations in experiments.

1.5.1 Infection phenotypes in *Drosophilidae*

Because of both the very recent description of most of the naturally occurring *Drosophila* virome, and the practical difficulties of viral isolation, only 8 (<10%) of the identified *Drosophila*-associated viruses have been isolated for experimental use. This had led to the use of non-native viruses in *Drosophila* experiments, such as Cricket paralysis virus, Sindbis virus (mosquito-vectored), and Vesicular stomatitis virus (mosquito-vectored), which, though they provide insight into the conserved insect antiviral immune response (see review Xu & Cherry 2014), provide little information on the fitness effects of insect viruses in the wild. The seven isolated viruses are Drosophila C virus (DCV) (Jousset *et al.* 1977), Drosophila A virus (DAV) (Christian 1987), *Drosophila melanogaster* sigma virus (DmelSV) (Fleuriet 1988), Drosophila melanogaster Nora virus (DmelNV) (Habayeb *et al.* 2006), *Drosophila innubila* Nudivirus (DiNV) (Unkless 2011), Kallithea virus (KV) (Palmer *et al.* 2018b), La Jolla virus (Carrau *et al.* 2018), and Galbut virus (Cross *et al.* 2020). Isolation is usually the first step to characterising fitness effects. Based on experimental data from these currently isolated viruses, their fitness costs to infected *Drosophila* range from the severe to the undetectable.

Fitness effects on the host are best characterised for viruses that induce severe pathology. For example, systemic infection with DCV (+ssRNA, *Dicistroviridae),* the most commonly used native *Drosophila* virus in experimental work, causes metabolic depression, reduced locomotory ability, and eventual mortality in Dmel adults (Arnold *et al.* 2013a). In addition, the two isolated DNA viruses of *Drosophila,* DiNV and KV, can both significantly reduce female fecundity (Unkless 2011; Palmer *et al.* 2018b). The vertically transmitted DmelSV (-ssRNA, *Rhabdoviridae*) causes little adult mortality but does incur a reduction in Dmel host fitness of ~25% (Yampolsky *et al.* 1999; Wilfert & Jiggins 2013). However, the stage of the *Drosophila* life cycle at which this reduction in fitness is incurred remains unclear. It is possible that this reduction in fitness is due to a 10-20% reduction in egg viability, as reported by two separate studies (Seecof 1964; Fleuriet 1981). Current data also suggests that persistent infection with DmelNV produces no obvious pathologies, despite infection producing the differential

expression of genes from several immune-related pathways (Cordes *et al.* 2013; Lopez *et al.* 2018). Overall, as only these 8 *Drosophila*-infecting viruses have been isolated, we have little idea of the full range of fitness consequences caused by the *Drosophila* virome.

Of the isolated *Drosophila* viruses, all but Kallithea virus (KV), and *Drosophila innubila* Nudivirus (DiNV) are RNA viruses, and all but DiNV have been isolated from Dmel (cell culture or live flies). Therefore, they represent only a small sample of the virome of the *Drosophila* genus and a biased sample of its diversity. Consequently, the pathology, transmission route, and fitness costs of most *Drosophila* viruses are still unknown. Galbut virus, a common partitivirus of Dmel, was recently isolated and confirmed to be replicating in Dmel gut and reproductive tissues (Cross *et al.* 2020), opening the way for its effect on host fitness to be characterised. With more such studies, a more accurate picture of the pathology of the whole *Drosophila* virome could be uncovered.

1.5.2 Transmission routes of *Drosophila* viruses

In the wild, an individual fruit fly may become infected with a virus in multiple ways; including ingestion of the shed virus from their environment (faecal oral transmission) as seen in DCV (Plus *et al.* 1975), or inheritance of the virus from a parent through gametes (vertical transmission) as seen in some sigmaviruses (Longdon *et al.* 2011), and Galbut virus (Cross *et al.* 2020). However, in the lab, most studies using isolated viruses use wounding, and injection into the thorax or abdomen as a method of systemic infection, an infection route which is likely to be extremely rare in the wild (see methods Merkling & van Rij 2015). Infection by wounding not only introduces the effects of injury into the experiment, but bypasses potential evacuation of the pathogen, and the defences of the epithelial gut barrier (Vodovar et al. 2005). As a result, some injected viruses have been shown to infect different tissues when compared to oral delivery. For example, faecal-oral delivery of DmelNV results in the infection of 5-15% of midgut cells, but injection into the body cavity also causes infection of the epidermis, cardia and occasionally the reproductive tract (Ekström & Hultmark 2016). This

tissue tropism based on infection route can also create different infection outcomes (DCV - Gupta *et al.* 2017; and see review - Mondotte & Saleh 2018). Additionally, the *Drosophila* anti-viral immune response can also be influenced by the method of transmission, demonstrated by the difference in Toll pathway responses to DCV and DmelNV on oral and systemic infection (Ferreira *et al.* 2014).

Other insect-pathogen interactions are also influenced by the route of transmission. Adaptation of *Drosophila* to bacterial infection is influenced by the route of transmission of the pathogen – be it oral or systemic injection (Martins *et al.* 2013). Congruently, across the insect phylogeny, the outcome of Deformed Wing Virus (DWV) infection in Honeybees is also influenced by transmission route, with blood-borne viral transmission by an ecto-parasitic mite inducing greater mortality than oral infection (Wilfert *et al.* 2016; Ramsey *et al.* 2019). An assessment of the fitness effects of insect viruses which utilises natural infection routes will give a more realistic estimate of their costs to wild populations.

1.6 Objective of the thesis

The overall aim of this thesis is to demonstrate the utility of the multi-host, multi-virus system of *Drosophilidae* and their associated viruses for studies of insect-virus community dynamics. By gathering data from this system, we can contextualise laboratory studies of anti-viral immune responses and co-evolution in insect-virus systems, and train predictive models of viral sharing, and host shifting.

As described in this introduction, concentrated surveys of multi-host, multi-pathogen systems can help us to understand how many viruses regularly infect, and co-infect insect communities, species, and individuals. To this end, in chapter two I describe the monthly sampling of a temperate community of sympatric *Drosophila* species, and their viruses, over three years. To better characterise the number of viruses infecting, and co-infecting, sympatric Drosophilids, I use metagenomic sequencing of these collections to describe known and new viruses present in this system. Few studies currently examine variation in the host-range, and

prevalence of insect viruses in multi-host communities (but see Hermanns *et al.* 2021). I use RT-PCR surveys to describe *Drosophila* RNA virus host range, and the individual-based prevalence of ten of these viruses over time and space.

In comparison to RNA viruses, DNA virus infections appear to be relatively rare in *Drosophila*. As a result, fewer *Drosophila*-infecting DNA viruses have been identified, and variation in their prevalence and genetic diversity remain undescribed. To begin to fill in this gap in our knowledge, in chapter three I use a European-wide dataset of pooled Dmel DNA to characterise spatiotemporal variation in the prevalence of three *Drosophila* DNA viruses, and to make initial measures of viral population genetic diversity.

Most studies of the fitness effects of insect viruses, and the resulting strength of virus-driven selection in insects, use injection of a virus isolate. This infection route is rare in the wild, and very few *Drosophila* viruses have been isolated for experimental use. In chapter four, I attempt to circumvent the need for isolated viruses when characterising these effects, and use contact to infect *Drosophila melanogaster* with a subset of their native viruses, and viruses carried by their sympatric sister species. To assess the strength of selection these viruses might impose on the host immune system I then measure the viral-induced reduction in offspring production and lifespan.

2 The discovery, host range and prevalence of RNA viruses infecting Scottish Drosophilids

Initial workflows for metagenomic virus discovery were written by Darren Obbard for previous studies, and DJO ran this initial workflow on the Sept-Oct '16 sequencing dataset. All wild collections, fly, lab work and other bioinformatics analyses were carried out by MW, after initial instruction was given by Nathan Medd, Ferghal Waldron and Darren Obbard.

Data availability

Data, shell scripts and code for figures in this chapter can be found in a GitHub repository at https://github.com/megan-a-wallace/RNA_virus_discovery_dynamics.

All virus genomes described in this chapter can be found in this repository, and on the Obbard lab website (https://obbard.bio.ed.ac.uk/data.html).

2.1 Introduction

Predicting the patterns and drivers of viral prevalence, within and between both species and populations is one of the most important topics in modern disease biology. Insect transmitted viruses are receiving increasing attention because of the global burden of arthropod-borne viral diseases (Folly *et al.* 2020), the potential of viruses as disease control agents (Carlson *et al.* 2006), and the destructive effects of viral diseases in economically important insects (eg. Wilfert *et al.* 2016). For vertebrate viruses, data on their prevalence in space and time, and their genetic diversity, can be combined to infer patterns of geographic spread (eg. Worobey *et al.* 2020), and host shifting. In comparison, the ecological and evolutionary dynamics of insect-virus systems, with the notable exception of honeybee viruses, in particular Deformed wing virus – reviewed in Grozinger & Flenniken (2019), remains relatively understudied. Because of the unique nature of the vertebrate adaptive immune response in comparison to all other animal phyla, this gives us a biased view of the co-evolutionary dynamics we expect

to see in host-pathogen systems. Based on co-evolutionary theory, we would expect insects and their viruses to be involved in reciprocal, adaptive genetic change (Woolhouse *et al.* 2002), with viruses driving selection for host resistance, and hosts driving reciprocal selection for repressors of this immune response. This view is supported by the rapid adaptive evolution seen on insect antiviral RNAi genes (Obbard *et al.* 2006, 2011; Kolaczkowski *et al.* 2011), and the presence of suppressors of this response in some virus genomes (Bronkhorst *et al.* 2014; van Mierlo *et al.* 2014). However, the consistency of this selection pressure is dependent on how specialised insect viruses are to their host species (host range), how often they switch between host species, and how consistently prevalent they are at a local scale.

There are few studies that examine the host range, and local-scale variation in the prevalence of wild insect viruses (but see Hermanns *et al.* 2021). Because of this, we don't have a clear picture of how host-specific or generalist insect viruses are, and how consistent the virosphere of wild insect species is across space and time. By quantifying these ecological parameters we could provide context to studies of insect-virus co-evolution, and find out why the evolution of insect immune genes seems to be able to compete with the evolution of viral virulence factors, despite the high mutation rates of viruses (see review Obbard & Dudas 2014). Ecological factors could play a key role in driving insect-virus co-evolutionary dynamics. For example, virus-driven selection on a rare host species' immune system could be dominated by spillover of the viruses from a sympatric, and more common host species. Or, host generalism could be favoured in a system where host species availability fluctuates, leading to fewer genotype by genotype interactions.

The model species *Drosophila melanogaster*, other sympatric *Drosophila* species, and their associated viruses provide an ideal model system in which to characterise the ecological context of insect-virus evolution. Numerous studies have characterised the anti-viral defence mechanisms of *Drosophila* (reviewed in Mussabekova *et al.* 2017), and metagenomic sequencing studies have identified >130 of their naturally-occurring viruses (eg. Brun *et al.* 1980; Webster *et al.* 2015, 2016; Medd *et al.* 2018). By conducting further, spatially

concentrated, and repeated metagenomic surveys of the sympatric community of U.K drosophilids, we can edge closer to a complete picture of the viral diversity present in this system.

Since 2010, studies of *Drosophila* viruses have used large scale metatranscriptomic sequencing to characterise the viral diversity associated with drosophilids, and to make a preliminary assessment of their host range (eg. Webster *et al.* 2015; Medd *et al.* 2018). However, within and between insect species, the amount of viral reads can be extremely variable, and read abundance across a pool of hosts is not necessarily predictive of prevalence (Batson *et al.* 2020). Barcode switching between sequencing samples run on the same lane can also lead to the miss-assignment of viruses to host species (Ballenghien *et al.* 2017). Because of this, only one published paper (Unkless 2011) has attempted to estimate *Drosophila* virus prevalence in any species other than *D.melanogaster* and *D.simulans*. To more accurately determine the prevalence, and host range of *Drosophila* viruses, we can use single host transcriptome sequencing or PCR assays of individuals and single species pools. Previous studies of naturally-occurring insect specific viruses have found variation in prevalence across host species (Shi *et al.* 2017), with season (Runckel *et al.* 2011) and across global locations (Webster *et al.* 2015), but we have little idea of the consistency of host-virus associations at smaller spatial scales. At a smaller spatial scale, differences in viral prevalence across host species could persist if they differ in their susceptibility to viral infection (Longdon *et al.* 2011a), or if the virus can block some host species immune defences better than others (eg. van Mierlo *et al.* 2014). Repeated sampling of *Drosophila* and their associated viruses, over time, and at a more condensed spatial scale than in previous studies, could help to make realistic estimates of the consistency of insect-virus associations.

These single host assays not only allow us to estimate virus prevalence, but also the rate of co-infection in wild insect communities. Viral co-infection can sometimes lead to viral replacement (Vazeille *et al.* 2016), or increased virulence of viral infections (Kuwata *et al.* 2015), but there are few estimates of its frequency in wild insects. In *Drosophila*, rates of co-

infection have previously been estimated at around 20% (Webster *et al.* 2015; Shi *et al.* 2018), but could be as high as 80% based on the rates recently ascribed to mosquitoes in California (Batson *et al.* 2020).

In this chapter, I use short read sequencing of multi-species pools to further characterise the diversity of RNA viruses associated with UK drosophilids. I identify new RNA viruses, and characterise the host range of these new viruses by RT-PCR. I also characterise the presence/absence of known viruses across host species in this system, and the individual-based prevalence of a subset of these viruses from single flies and species-site pools. Together, these data expand the number of recorded Drosophila viruses by 10%, demonstrate the utility of this wild system for studies of multi-host, multi-virus community dynamics, and contribute significant data for further studies of the determinants of viral host range, and virus prevalence in insects.

2.2 Methods

2.2.1 Collection and identification of multiple species of Drosophila

Between September 2016 and October 2018 I collected a total of 2227 wild *Drosophila* from 20 field sites within a 20km^2 area encompassing the City of Edinburgh, along with parts of East and Mid-Lothian. Field sites were selected for their proximity to both urban and woodland or rural environments, due to the *Drosophilidae*' s liking for woodland, anthropogenic refuse, and fruit farms. See Fig. 2.1 for a map of the sampling sites, and Table S2 1 for further site details, including their coordinates. Every three months, sampling sites were randomly selected from location-based site groups for each month, and sampled without replacement to prevent sampling at the same site two months in a row. Monthly, I placed one bottle and one bucket trap, both with banana and yeast bait, at five locations across the sampling area for five days, until May 2017, after which I increased the number of monthly sampling sites to six. Traps were laid monthly over the 2016-2017 winter, but in 2017 and 2018, collections were ceased for the winter months after the first zero-catch month. I started spring sampling after the first collections of *Drosophila* from a pilot trap in Kings Buildings, Edinburgh (55.922319, -3.172339).

On the same day as collections, I identified collected flies to species by morphology and pooled the flies into pots of 1-30 individuals by species and sampling site, ensuring a range of pot sizes for each species-by-site combination for later prevalence calculations. For each of the 737 pooled *Drosophila* samples collected, I then used a manual Phenol-Chloroform protocol to isolate total RNA. The flies were macerated in TRIzol reagent (Invitrogen, 100-750µL depending on the number of flies in the pool), RNA extracted using the manufacturer's instructions and the final pellet re-suspended in 30 µL RNAase free H_2O.

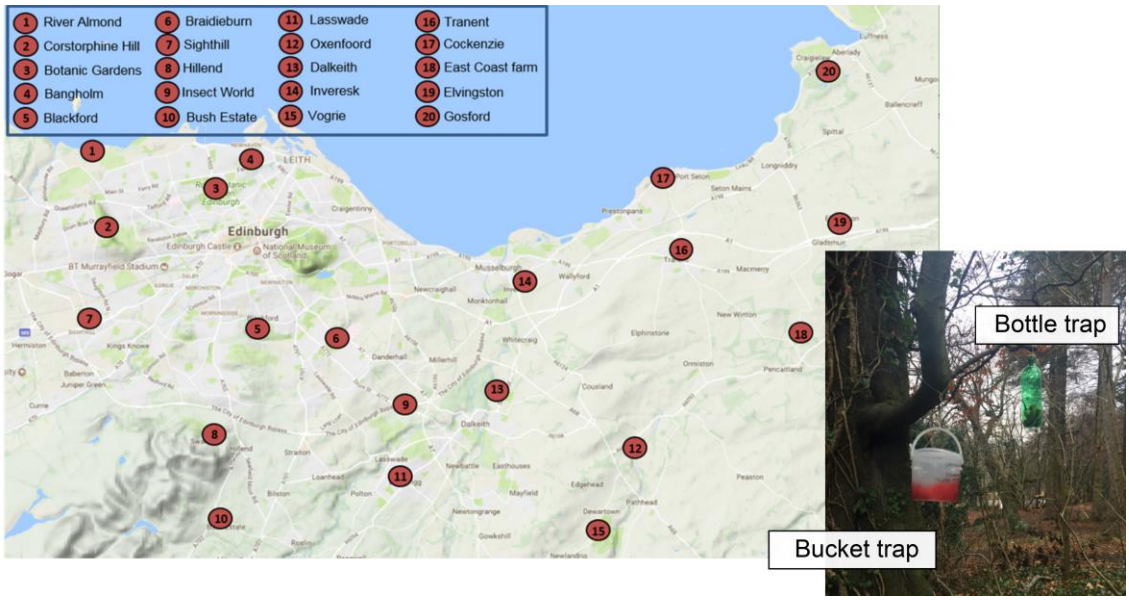


Fig. 2.1. Sampling strategy for Drosophila collections. The plot shows a map of the sampling sites used for collections, and sample site names. The inset panel shows an example of the banana and yeast traps baited traps used at each location, which were sampled from using aspiration and netting. The landowner's permission was obtained for each site.

2.2.2 Time structured sequencing of mixed-species pools of Drosophilidae

To characterise the viruses infecting Drosophila in the sampling area, and, in particular, to identify any new drosophilid-infecting viruses, I used RNA sequencing of large temporally structured pools of flies. I separated the collected *Drosophila* into five multi-species pools by date of collection: Sept-Oct 2016, Dec 2016-June 2017, July-Oct 2017, Dec 2017-June 2018 and July-Oct 2018. I separated the collections like this to capture the potential difference between viruses present in periods of low and high host density (eg. December-June vs. July-October). The quality of extracted RNA was checked using gel electrophoresis, and a Nanodrop spectrophotometer and Invitrogen Qubit Fluorometer assays used to measure RNA concentration, and check levels of any DNA or phenol contamination. For each sequencing pool, I pooled the samples by species, and then pooled the species so that the per-fly mass of RNA was equal. However, to maximise the chance of viral discovery for the September-

October 2016 pool, as the majority of flies were *D.immigrans*, the sequencing pool comprised 50% RNA mass of from *D.immigrans*, and 50% from all other species combined with equal per-fly mass of RNA. I submitted the five pools of RNA to Edinburgh Genomics for five TruSeq stranded total RNA-seq library preparations, with no poly-A selection (as many viruses don't have poly-A tails), and RiboZero Gold treatment to deplete ribosomal RNA (Benes *et al.* 2011). Each library was then sequenced on an Illumina HiSeq 4000, producing 150nt (Sept – Oct '16) or 75nt (all other libraries) paired-end Illumina reads.

I *de novo* assembled raw reads from the total RNA sequencing into contigs using Trinity (Grabherr *et al.* 2011), whilst using Trimmomatic (Bolger *et al.* 2014) with the default parameters to quality trim at ends. I then clustered the hits to thin results using cd-hit-est (Fu *et al.* 2012) at a divergence level of 0.85, as this is the divergence often used to delimit virus 'species'. To create a set of query proteins from which to identify viruses, I re-formatted the assembled and filtered contigs using the FASTX-Toolkit (Hannon 2010), and identified the longest open reading frame (ORF) from each. I then called all translations and concatenated the proteins using the seqinr (Charif & Lobry 2007) package in R version 3.6.1 (R Core Team 2019).

2.2.3 Identifying virus-like contigs

For each dataset, I identified the likely sources of the assembled proteins by doing a DIAMOND (Buchfink *et al.* 2014) BLASTp (BLAST 2008) search using the concatenated protein sequences found in the sequencing runs as queries, and an e-value of 0.01. I searched against a protein database that included the most common fly sequencing contaminants (eg. bacteria, flies, nematodes, protozoa) and all non-redundant viral proteins in NCBI databases (accessed by filtering for virus general identifiers in an NCBI taxdump download), and then selected matches to this protein database which had accession numbers of viruses. This workflow produced a set of virus-like contigs for each dataset, from which I identified known and new viruses.

To identify known viruses I used the identified virus-like nucleotide sequences from each of the five datasets to search against a database of *Drosophila* viruses identified by our lab and others (correct as of November 2019 - see https://obbard.bio.ed.ac.uk for up to date list of sequences) using BLASTn and DIAMOND. To select only high quality matches to known viruses, I filtered for blast hits with an e-value smaller than 1e-50, and matches with at least 90% similarity at the nucleotide level. I manually inspected any contigs matching to known *Drosophila* viruses where the matching contigs were longer or more complete than the current genome assembly, looking for the most credible protein coding regions, and saving translations. I manually inspected the virus-like contigs not identified as known *Drosophila* viruses and identified the most credible looking novel viruses from them. I considered putative viruses more credible if; 1) they were >2kb in length, 2) BLASTp could identify reasonable protein translations from the ORFs in the virus contig, and 3) the viruses came from a family of known invertebrate viruses. I also manually checked for contaminants and fly genomic material in the putative contigs. I then used targeted BLASTn searches to identify if the putative new viruses were present in multiple datasets. For each virus, if it was found in multiple datasets I saved the assembly with the most complete genome for further analyses.

2.2.4 Read mapping and viral genome organisation

2.2.4.1 Mapping reads

To verify that my new virus assemblies were structured correctly, without gaps, I mapped the raw paired-end reads from each of the datasets back to a database containing the putative new virus genomes using bowtie2 (Langmead & Salzberg 2012), always using the –very-sensitive option to reduce cross-mapping between closely related viruses. I measured coverage across the genomes with Bedtools (Quinlan & Hall 2010), and plotted this with the ggplot2 package (Ginestet 2011) in R. To confirm the putative strandedness of the new virus genomes, and as an initial check for replication, I checked the number of reads mapped to each strand separately using samtools (Li et al. 2009) and re-checked the orientation of any

viruses with unexpected mapping patterns. All new virus sequences can be found in the *Drosophila* virus list on the Obbard lab website (https://obbard.bio.ed.ac.uk/data.html), along with the genomes of eight known *Drosophila* viruses that I extended. I provisionally named the viruses after sampling sites, or locations near sampling sites, in my study area.

With this final list of new viruses, I mapped reads from each dataset to a database of all new and known *Drosophila* viruses and Cytochrome oxidase 1 (COI) genes from each of the *Drosophila* species, using bowtie2. To get a measure of the relative abundance of viral transcripts in comparison to those from the host, I counted reads mapping to each virus, normalised these counts by virus genome length, and the reads mapped to all *Drosophila* species COI genes. I plotted these results in a heatmap produced using the R package pheatmap (Kolde 2015).

2.2.4.2 Genome organisation and viral phylogenies

I then identified likely protein coding regions, inferred functionality of proteins, and visualised viral genome organisation for the new, and extended known viruses. To do this I used a combination of manual ORF identification in Bioedit (Hall 1999), BLASTp searches of translations and the VGAS viral genome annotation system (Zhang *et al.* 2019a) to identify ORFs and annotate possible protein functions. I then used the R package gggenes (Wilkins & Kurtz 2020) to plot viral genome organisation. To display the evolutionary relationships between the new *Drosophila* viruses and other viruses and virus-like sequences, I built small phylogenies from their polymerase proteins. For *Drosophila* Burdiehouse burn Chuvirus I couldn't identify the replicase protein sequence, and so I used the longest ORF. I created a multiple sequence alignment of the top ~5 BLASTp hits, plus a representative of each surrounding genus in the virus family (Walker *et al.* 2020) using Clustal omega (Sievers *et al.* 2011), and default parameters for protein sequences. Within the IQ-TREE software (Nguyen *et al.* 2015), I used ModelFinder (Kalyaanamoorthy *et al.* 2017) to select the best fit amino acid substitution model for the viral phylogenies, then constructed maximum likelihood trees,

and calculated branch support using ultrafast bootstrap approximation (Hoang *et al.* 2018). Trees were mid-point rooted, and any downstream annotation was completed using the ggtree package (Yu 2020) in R.

2.2.5 *Characterising the host range of new and known Drosophila viruses*

2.2.5.1 Primer design and PCR assays

I used the Primer3 (Untergasser *et al.* 2012) primer design tool in Geneious 10.1.3 (http://www.geneious.com/) to design RT-PCR primers for the new viruses (Table S2 ***2***), and test that they were detectable in the sequencing pools. I designed the primers such that they shouldn't cross prime to known closely related *Drosophila* viruses but do target relatively conserved protein regions of the viral genome. I tested these primers on the pools of collected flies for sequencing. To do this, I synthesised cDNA using random hexamer primers (4µL 10 µM random primers per 1µL total RNA) which were incubated with the total RNA from each sequencing pool (70ºC, 5 minutes). I then added RNase free H_2O, 10 µM mixed dNTPs, Moloney Murine Leukaemia Virus (M-MLV) reverse transcriptase (200 units/µl) and 5x M-MVL reaction buffer to the 5µL RNA-random primer mix and incubated at 37ºC for 60 minutes. All PCR assays are expected to work with a master mix of the following volumes per 1µl of cDNA template - 1µl 10xNH4 buffer, 0.25µl 50mM MgCl2, 0.05µl 5U/µl BIOtaq DNA polymerase, 0.3µl 10mM mixed dNTPs, 7.5µl RNase free H2O, and 0.5µl of both 10µM forward and reverse primers. PCR assays should then be run on a thermocycling regime of 94ºC for 5 minutes, then 5-10 x (94ºC for 15s, * ºC for 30s, 72 ºC for 1 minute) dropping 1^0C every cycle, then 25-30 x (94ºC for 15s, * ºC for 30s, 72 ºC for 1 minute), then 72 ºC for 5 minutes, with the * ºC adjusted for virus assays. Presence or absence of a virus was determined using gel electrophoresis, stained using GelRed 10,000x, on a 2% agarose gel at 90V. Details of the newly designed working RT-PCR primers, with the Tm for each primer pair can be found in Table S2 2.

2.2.5.2 Host species COI sequencing

To confirm that the morphological species identifications were correct for rarer species (not *D.immigrans*, *D.melanogaster*, or the *D.obscura* group), I verified them via Sanger sequencing of a region of the Cytochrome oxidase 1 (COI) gene (primers at 10 μM = F – TTCAACAAAYCATAARGAYATTGG, R - CTTCAGGRTGNCCAAARAATCA). COI PCR products were cleaned using Exo-SAP-IT ® Express Reagent (Applied Biosystems) following manufacturer's instructions, before the BigDye ® Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems), was used for sequencing reactions, with 3.2 μM primers. Capillary electrophoresis was performed at Edinburgh Genomics, and the resulting sequences manually trimmed at ends using FinchTV (Geospiza 2004) when quality scores dropped below ~ 80. Geneious v10.1.3 (http://www.geneious.com/) pairwise alignment software was used to generate consensus sequences from the forward and reverse reads, where the base call with the highest quality score took precedence if there was a mismatch. A BLAST search against NCBI databases (Altschul 1990) was then used to confirm species identity. To summarise the diversity of the host species community, I used these confirmed species identifications to calculate Shannon's diversity index using the R package vegan (Oksanen *et al.* 2012), which accounts for both abundance and evenness of the species present.

2.2.5.3 Host range screens

For each of the newly described viruses, I then characterised their host range in this system. I used the newly designed primers to screen species pools from each sequencing timeframe for presence/absence of the virus. For segmented viruses, I assayed each sequencing pool and species for all segments separately, and used the co-occurrence of segments across species-by-pool combinations to assay the support for the assignment of segments to the same virus. To assay the host range of a larger portion of the virosphere in this system, I combined this data with RT-PCR surveys for 37 additional previously-published *Drosophila* viruses (Webster *et al.* 2015, 2016; Medd *et al.* 2018a) across species pools. For the four most

common species (*D.immigrans*, *D.obscura*, *D.subobscura*, and *D.subsilvestris*), these pools were divided into the years of sampling. All reverse transcription reactions and PCR assays were done using the methods described above, and using the primers and conditions in Table S2 3. I created heatmaps displaying the presence/absence of both new, and known viruses across species using the R package pheatmap *(Kolde 2015).*

2.2.6 Characterising variation in Drosophila virus prevalence and sequence across host species, space, and time

To quantify the variation in virus prevalence across host species, time, and site, I assayed the 737 samples (pools or individual flies of each species x site x month combination) individually for ten viruses. Individual-based prevalence assays were done for Galbut virus (and its satellite, Chaq), three Nora viruses (Drosophila immigrans Nora virus, Drosophila subobscura Nora virus, and Drosophila melanogaster Nora virus), the closely related Prestney burn virus, Motts mill virus and Grom virus, newly described Tranent virus, Muthill virus, and Drosophila immigrans sigma virus (first described in van Mierlo *et al.* 2014; Webster *et al.* 2015, 2016; Longdon *et al.* 2017). All reverse transcription reactions and PCR assays were done using the methods described for host range assays, and using the primers in the Table S2 ***3***.

2.2.6.1 Sequence diversity

To check for cross-priming between closely related viruses and, for the multi-host viruses, to look at genetic differentiation between viruses in different hosts, I Sanger sequenced the PCR products from five viruses, in a selection of individual flies and small pools. For the five viruses (Prestney burn, Grom, Motts mill, Drosophila immigrans Nora (DimmNV) and Drosophila immigrans sigma virus (DimmSV)), I used primers that target regions of protein 1 (Motts Mill, Grom and Prestney burn), or the polymerase (DimmNV and DimmSV). For DimmSV I also sequenced a region of the N (nucleocapsid) gene. The primers used are highlighted in Table S2 ***3***. I used Sanger sequencing methods described above and, additionally, removed sequences with heterozygous sites which could signal that multiple, genetically divergent viral

genotypes infected the sample. I checked that the sequences were from the target virus by comparison to NCBI databases (Altschul 1990), and then aligned them using a codon model in PRANK multiple sequence aligner (Löytynoja 2014). These alignments were used to construct phylogenies, whilst inferring the most likely host species on ancestral nodes, in BEAST v1.10.4 (Rambaut *et al.* 2018). I used the SDR06 substitution model (Shapiro *et al.* 2006), and default priors, other than a lognormal prior with a mean and initial value of $5x10^{-4}$, and standard deviation $2x10^{-4}$ (a reasonable rate for RNA viruses), on the strict clock rate. It's likely that this prior heavily influenced the time calibration of these trees, as I have a very limited spread of tip dates over time. Whether a constant or exponential population size tree coalescent was more appropriate was determined from the estimates of population growth rate in preliminary models. Preliminary trees, where the rate of host species switching on the ancestral nodes wasn't fixed, showed there wasn't enough power to make rate estimates for specific species pairs. So, for all viruses apart from the Drosophila immigrans sigma virus, I didn't allow transition rates between host species to vary independently, such that a single mean transition rate was estimated. For DimmSV I used the alignments from both regions of the virus genome with linked tree topologies, but unlinked substitution rate and clock models, to infer ancestral host species states. For all phylogenies, two Markov chain Monte Carlo (MCMC) chains with a length $5x10^7$ were run, sampling 10,000 trees on each chain. I examined trace files for sufficient effective sample size and chain convergence before combining runs and generating a maximum clade credibility (MCC) tree in TreeAnnotator v1.10.4 (Drummond & Rambaut 2007), setting the burn-in to 10% of the total MCMC chain length. The MCC trees were viewed in FigTree v1.4.4 (http://tree.bio.ed.ac.uk/) and using the R package ggtree (Yu 2020).

2.2.6.2 Variation in prevalence

I then inferred the underlying prevalence of each of the ten viruses from their presence/absence across pools, and single flies. To do this, I used a maximum-likelihood function based on a model of pooled Bernoulli trials (eg. Speybroeck *et al.* 2012). This function

searches across a range of prevalence values to identify the prevalence which maximises the likelihood of the observed number of pools/single flies testing positive, given the sizes and number of pools. To ask whether prevalence varied significantly across host species, collection seasons (early and late: pre-July or July onwards), and sampling sites, I compared the log likelihood of prevalence when it was assumed to be the same across all combinations of these covariates, and the summed log likelihood when it was assumed to vary. I selected the best model for each virus using Akaike weights, which represent the relative likelihood of a model compared to all alternative models (calculated from the change in Akaike Information Criterion (AIC)). I compared four models, 1) no covariates (uniform prevalence), 2) prevalence ~ species, 3) prevalence ~ species + season, and 4) prevalence ~ species + season + site.

2.2.6.3 Co-infection

I then inspected patterns of viral co-infection, assessing if any viral combinations occurred significantly more or less than expected. I analysed the frequency of each viral combination across all species, and then within each species. The probability that the two viruses co-occur at the observed frequency of co-occurrence is derived from calculating the number of ways in which the two viruses could co-occur across the individuals present, given their individual frequencies, assuming a random and independent distribution of each virus (Veech 2013). I assessed the significance of these probabilities using the R package cooccur (Griffith *et al.* 2016), determining that the association was significant if the two viruses had a <0.05 probability of co-occurring at the observed frequency if they were randomly and independently distributed.

2.3 Results

2.3.1 Monthly collections of 15 Drosophila species over 3 years

Between September 2016 and October 2018, I was able to collect *Drosophila* at all 20 of my sampling sites across Edinburgh and the Lothians. Fly abundance varied between sites, from 0.09% (2 flies - Corstorphine hill) to 22.7% (505 flies - Blackford) of the 2227 total flies collected (Fig. 2.2 panel A). I collected the vast majority (79.9% of 2017 and 2018 collections) of *Drosophila* between the months of July and October (Fig. 2.2 panel B), showing that the population fluctuates with season (Basden 1954). I collected 15 species of the *Drosophilidae*. These species were *D.immigrans*, *D.funebris*, *D.hydei*, *D.busckii*, and *D.phalerata* (genus *Drosophila*, sub-genus *Drosophila*), *D.melanogaster*, *D.obscura*, *D.subobscura*, *D.subsilvestris*, *D.tristis*, and *D.helvetica* (genus *Drosophila*, sub-genus *Sophophora*), and *Chymomyza costata*, *Scaptodrosophila deflexa*, and *Hirtodrosophila cameraria* (out-groups to the *Drosophila* genus). A member of the *Drosophila virilis* species group was also identified from its COI barcode sequence but could not be reliably identified to species. The most commonly collected species was *D.immigran*s (59.9% of total flies), followed by *D.subobscura* (12.8%), *D.obscura* (8.85%), D.subsilvestris (7.36%), and D.melanogaster (7.23%). All other species combined made up less than 4% of the collections. There was a distinct shift in the species composition of collections between the early and late sampling seasons, with collections before July dominated by species of the Obscura group (D.obscura, D.subobscura, D.subsilvestris, and D.tristis), which made up 71.1% of collections, and later collections by. D.immigrans, which made up 67.1% of collections. The host species biodiversity was higher in the early sampling season (Shannon diversity index (H) = 1.573, effective number of species = 4.82), compared to the later sampling season (H = 1.202, effective number of species = 3.32).

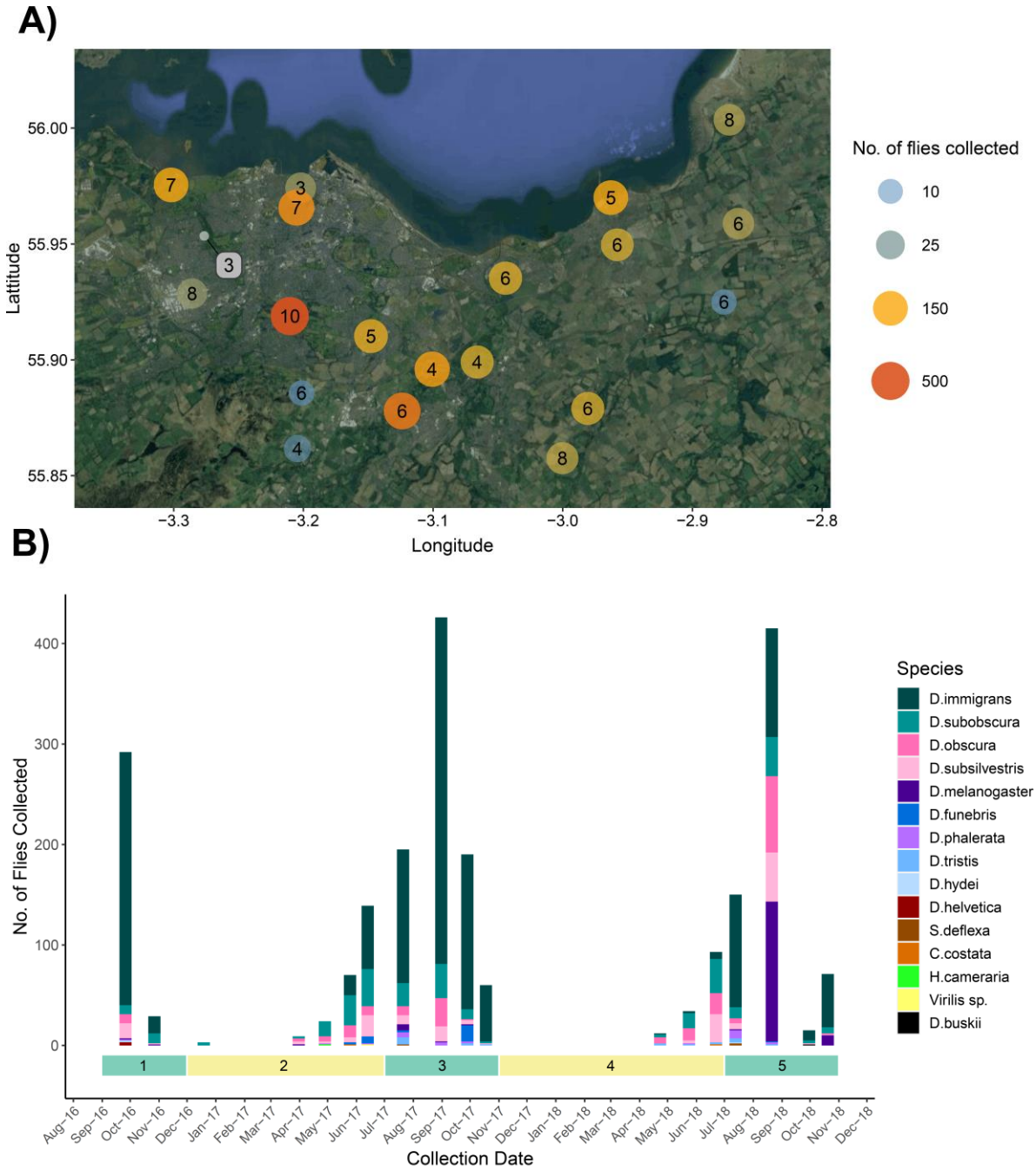


Fig. 2.2. Drosophila collection details. A) Sampling area with collection site annotated, points are coloured and sized by the number of flies collected at each location, and numbers indicate sampling effort (the number of times traps were placed) at each site across the 26 months of sampling. B) The plot shows the number of flies, with species as stacked bars, collected across the sampling period. Green and yellow boxes underneath the bars indicate the months of sampling which were pooled for five RNA sequencing runs.

2.3.2 Identification of known and new virus contigs, and their abundance across datasets

From the five metagenomic sequencing runs (1-5 in Fig. 2.2 panel B), I generated 324 million read pairs, ranging from 99 million (September - October 2016) to 44 million pairs (July – October 2017) per library. The reads were assembled into 180,000 (December 2016 – June 2017) to 45,500 (July – October 2017) putative contigs, from which I identified 17 new RNA viruses associated with Scottish Drosophilids (details of all new viruses in Table 2.1). The new viruses encompass negative sense single-stranded RNA viruses, positive sense single-stranded RNA viruses, and double-stranded RNA viruses, and eight are likely to be complete or near-complete genomes. I also identified a variant of Drosophila C virus (Jousset *et al.* 1977) in one of the sequencing pools (September-October 2016). At 92% capsid protein identity, this is not sufficiently divergent to warrant consideration as a separate virus, according to species demarcation criteria for the genus *Cripavirus* (Walker *et al.* 2020). Viral genome organisation, with open reading frames, and putative functional annotations is displayed in Fig. 2.3, and read coverage across the new viral genomes in Fig. S2. ***1***. The new viruses are from the orders *Jingchuvirales* (2), *Ghabrivirales* (2), *Martellivirales* (3)*, Reovirales* (2), *Wolframvirales* (1), *Mononegavirales* (2), *Nodamuvirales* (2), and *Bunyavirales* (4). Their phylogenetic relationships to other known viruses and virus-like sequences are displayed in Fig. 2.4.

From the five sequencing pools I was also able to improve the assemblies of eight already published *Drosophila* viruses. Each of these assemblies either extended the current genome, or filled in gaps in the published assembly. More details on each of these virus assemblies can be found in Table S2 **4**, and viral genome organisation is displayed in Fig. S2. 2. Of note, I completed the genome of two Rhabdoviruses previously described by their polymerases only - Drosophila tristis sigma virus (Longdon *et al.* 2015b) and Withyham virus (Webster *et al.* 2016) - and completed the genome of a Flavivirus, Hermitage virus, which was previously described in two separate contigs in Webster *et al.* (2016). The unclassified viral clade

containing Hermitage virus is thought to be basal to the *Flaviviridae* family (Webster *et al.* 2016), and also contains another *Drosophila* virus, *Takaungu* virus (53.19% amino acid identity on the polyprotein) assembled from mixed-Drosophilid reads from Kenya (Webster *et al.* 2015).

Mapping non-rRNA reads from all sequencing libraries to a database of new and known *Drosophila* viruses detected a total of fifty six RNA viruses present in this system. The number of virus transcripts relative to host COI transcripts are displayed as a heatmap in Fig. 2.5. The percentage of reads mapping to *Drosophila* viruses varied from 0.967% in September-October 2016, to 8.9% in April-Jun 2018, with an average of 3.74% of reads being viral per dataset. Forty viruses had an average of >= 0.01 viral copies relative to host COI. The virus with the highest average relative number of viral transcripts (6.71) was Prestney burn virus (Webster *et al.* 2016), a positive sense single-stranded RNA virus originally described in *D.subobscura*, and closely related to two other Drosophila viruses, Grom virus and Motts Mill virus (cf. Polerovirus/ sobemovirus) (Webster *et al.* 2015, 2016). The viruses with the next highest relative numbers of viral transcripts - Drosophila immigrans Nora virus (van Mierlo *et al.* 2014), Craigie's Hill virus (Webster *et al.* 2015), Larkfield virus (Medd *et al.* 2018a), and the Cockenzie Drosophila C virus variant - all show viral transcript numbers greater than host COI. Of note, I detected no DNA viruses of *Drosophila* in any of the sequencing pools.

It's likely that some of the reads in the September – October 2016 pool, in particular those mapping to viruses identified in D.suzukii in Medd *et al.* (2018), are the result of barcode switching (Ballenghien *et al.* 2017) between this library, a D.suzukii library, and D.melanogaster laboratory fly library sequenced on the same Illumina lane. This is likely the reason for the different assemblage of viruses present in this dataset, including Mogami virus, Saiwaicho virus (Medd *et al.* 2018a) and Drosophila A virus (Ambrose *et al.* 2009), which are absent in all other datasets. I did not find Drosophila A virus infections in the host range RT-PCR screen, and so reads mapping to this virus are confirmed as barcode switching. In the other sequencing pools, there is a notable lack of viruses most commonly used in *Drosophila*

927 laboratory experiments, such as Drosophila A virus (Ambrose *et al.* 2009), Drosophila C virus
928 (Jousset *et al.* 1977), and D.melanogaster sigma virus (Fleuriet 1988). However, the absence
929 of D.melanogaster sigma virus, which is vertically transmitted, could be due to the
930 inconsistency of Dmel presence in this system.

Provisional name	Classification (order, family)	Sequencing pool	Description
Drosophila Burdiehouse burn Chuvirus	*Jingchuvirales, Chuviridae*	Sep – Oct ‘16	Incomplete genome of a –ssRNA chu-like virus. Related to *Lampyris noctiluca* chuvirus-like virus 1 (unpublished - MH620819.1), Wuchang Cockroach Virus 3 (Li *et al.* 2015b), Coleopteran chu-related virus OKIAV151 (Käfer *et al.* 2019) and other chu-like viruses of invertebrates. [4.8 kbp fragment encoding three proteins]
Drosophila Cockenzie Picornavirus (DCV variant)	*Picornavirales, Dicistroviridae*	Sep – Oct ‘16	Whole genome of a +ssRNA virus, most likely a variant of *Drosophila* C virus (DCV) (Jousset *et al.* 1977), as the putative capsid polyprotein is 92% identical to DCV (NP_044946.1). [9.2 kbp fragment encoding two proteins]
Drosophila Craighall Totivirus	*Ghabrivirales, Totiviridae*	Apr – Jun ‘18	Near-complete genome of a dsRNA Totivirus. Most closely related to Inverleith virus (82.2% genome-wide nucleotide identity), see below. More distantly related to Hubei toti-like virus 6 (Shi *et al.* 2016a), Diatom colony associated dsRNA virus 17 (Urayama *et al.* 2016) and a putative toti-like virus described from Dmel cell culture (SRR1197466 – polymerases ~35.2% identical) (Webster *et al.* 2015) [5.9 kbp fragment encoding two proteins]
Drosophila Crammond Virga-like virus	*Martellivirales, Kitaviridae*	Dec ’16 – Jun ‘17	Near-complete genome of a +ssRNA virga-like virus. Related to Hubei virga-like viruses 9 and 10 (Shi *et al.* 2016a), Abisko virus of moths (de Miranda *et al.* 2017), *Culex pipiens* associated Tunisia virus (Bigot *et al.* 2018) and other virga-like viruses of Diptera. [10.9 kbp fragment encoding five proteins]
Drosophila Dalkeith Chuvirus	*Jingchuvirales, Chuviridae*	Dec ’16 – Jun ‘17	Incomplete genome of a –ssRNA chu-like virus. Related to Shyang fly virus 1 (Li *et al.* 2015b), and Mogami virus (Medd *et al.* 2018a) identified in *D. suzukii* collected in Japan. [8.3 kbp fragment encoding two proteins]
Drosophila Glencorse burn Reovirus	*Reovirales, Reoviridae*	Jul – Oct ‘17	Five segments of a dsRNA reovirus. Related to Hubei odonate virus 15 (Shi *et al.* 2016a), *Anopheles hinesorum* orbivirus (Colmant *et al.* 2017a), and other orb-like reoviruses of insects. [Segment 1 – 3.8 kbp fragment encoding one protein. Segment 2 - 2.8 kbp fragment

			encoding one protein. Segment 3 – 2.9 kbp fragment encoding one protein. Segment 4 - 1.1 kbp fragment encoding one protein. Segment 5 – 1.1 kbp fragment encoding one protein]
Drosophila Gosford Narnavirus	Wolframvirales, *Narnaviridae*	Jul – Oct ‘18	Complete genome of a +ssRNA narnavirus. Related to Wilkie narna-like virus 2 (Shi *et al.* 2017), described in mosquitoes, Behai narna-like virus 22 (Shi *et al.* 2016a), described in Penaeid shrimp, and Plasmopara viticola lesion associated narnavirus 14 (Chiapello *et al.* 2020). [3.2 kbp fragment encoding one protein. Another possible 101aa protein encoded on the reverse strand, as described in Dinan *et al.* (2020).]
Drosophila Hillwood park Negevirus	*Martellivirales, Kitaviridae*	Apr-Jun ‘18	Near-complete genome of a +ssRNA nege-like virus. Related to Cordoba virus (Nunes *et al.* 2017) and Negev virus (Fujita *et al.* 2017) of mosquitoes. More distantly related to River almond virus (polymerases 24.5% identical). [9.2 kbp fragment encoding three proteins]
Drosophila Inveresk Nyamivirus	*Mononegavirales, Nyamiviridae*	Sep - Oct '16	Incomplete genome of a -ssRNA virus. Related to Soybean cyst nematode nyami-like virus (Ruark *et al.* 2018), Soybean cyst nematode socyvirus (Bekal *et al.* 2011), and Hymenopteran orino-related virus OKIAV85 (Käfer *et al.* 2019). [6.6 kbp fragment encoding one protein]
Drosophila Inverleith Totivirus	*Ghabrivirales, Totiviridae*	Apr – Jun ‘18	Complete genome of a dsRNA totivirus. Most closely related to Craighall virus, see above. More distantly related to Hubei toti-like virus 6 (Shi *et al.* 2016a), Diatom colony associated dsRNA virus 17 (Urayama *et al.* 2016), and a putative toti-like virus described from Dmel cell culture (SRR1197466 – polymerases ~36.4% identical) (Webster *et al.* 2015) [6.7 kbp fragment encoding two proteins]
Drosophila Lasswade Rhabdovirus	*Mononegavirales, Rhabdoviridae*	Sep – Oct ‘16	Incomplete genome of an –ssRNA virus. Related to Soybean thrip rhabdo-like virus 1 and 2 (Thekke-Veetil *et al.* 2020), Hymenopteran almendra-related virus OKIAV1 (Käfer *et al.* 2019), and vesicular stomatitis Cocal virus (Pauszek *et al.* 2008). [5.1 kbp segment encoding one protein]

Drosophila Midmar Tombusvirus	*Nodamuvirales, Sinhaliviridae*	Jul – Oct ‘17	Two segments of a +ssRNA tombus-like virus. Most closely related to Dansoman virus, a virus first described in *D. melanogaster* (Webster *et al.* 2015). More distantly related to Hubei tombus-like virus 42 described in flies (Shi *et al.* 2016a), and Chronic bee paralysis virus of honeybees (Olivier *et al.* 2008). [Segment 1 – 3.7 kbp fragment encoding two proteins. Segment 2 – 2.1 kbp fragment encoding one protein]
Drosophila North esk Phasmavirus	Bunyavirales, *Phasmaviridae*	Jul – Oct ‘18	Near-complete genome of a –ssRNA phasma-like virus, with three segments. Related to Hubei diptera virus 7 and 8 (Shi *et al.* 2016a), Apis Bunyavirus 2 (Remnant 2017), and other phasmaviruses of diptera. [Segment 1 – 6.6 kbp fragment encoding one protein. Segment 2 – 2.5 kbp fragment encoding one protein. Segment 3 – 2.2 kbp fragment encoding three proteins]
Drosophila River Almond Negevirus	*Martellivirales, Kitaviridae*	Apr – Jun ’18	Near-complete genome of a +ssRNA nege-like virus. Related to Daeseongdong virus 1 (Unpublished - MT096524.1), Castlerea virus (O’Brien *et al.* 2017) and Piura virus (Nunes *et al.* 2017) of mosquitoes, and also to other Negeviruses of insects. More distantly related to Hillwood park virus (polymerase proteins 24.5% identical). [9.6 kbp fragment encoding three proteins]
Drosophila Sighthill Phlebovirus	*Bunyavirales, Phenuiviridae*	Sep - Oct '16	L and M segments of a -ssRNA virus. Related to Grand Arbaud virus and Uukuniemi virus, both discovered in ticks (Palacios *et al.* 2013). Also distantly related to other phlebovirus-like contigs of ticks. Glycoproteins show similarity to those of Huangpi Tick Virus 2 (Li *et al.* 2015b). [6.9 kbp L segment encoding one protein + 3.3 kbp M segment encoding two proteins]
Drosophila Sunshine Bunyavirus	*Bunyavirales, Phenuivirdae*	Sep - Oct '16	L segment of a -ssRNA virus. Part of the Phlebo-bunya virus like clade defined in Shi *et al.* (2016). Related to Jiangxia Mosquito Virus 1 (Li *et al.* 2015b), and Kristianstad virus (Pettersson *et al.* 2019) in Culex mosquitoes. This virus was also identified infecting *D. immigrans* in Canada by Matt Ballinger, who gave this virus its name (unpublished, personal communication). [4.8 kbp fragment encoding one protein]

Drosophila Tranent Phlebovirus	*Bunyavirales, Phenuivirdae*	Dec ’16 – Jun ‘17	L, M and S segments, likely the complete genome, of a –ssRNA virus. Related to Hubei diptera virus 3 (Shi *et al.* 2016a), Santiago bunya-like virus (Mahar *et al.* 2020), and other Phlebo-like viruses of diptera. [6.8 kbp L segment encoding one protein, 4 kbp M segment encoding one protein, and a 1.8 kbp S segment encoding one protein]
Drosophila Vogrie Reovirus	*Reovirales, Reoviridae*	Sept - Oct '16	Three segments of a dsRNA virus. Related to the Phytoreovirus and Rotavirus genera, in particular, Rice gall dwarf virus (Moriyasu *et al.* 2007), Homalodisca vitripennis Reovirus (Stenger *et al.* 2010) and Hubei reo-like virus 11 (Shi *et al.* 2016a). [Segment 1 – 4.2 kbp fragment encoding one protein, Segment 2 – 4 kbp fragment encoding one protein, putative Segment 5 - 2.5 kbp fragment encoding one protein]

931 **Table 2.1. New viruses reported in this chapter.** Closest relatives based on BLAST hits to polymerase proteins where available, and up to
932 date as of February 2021. Classification with reference to the 2019 release of the ICTV report. Sequencing pool refers to the pool from which
933 reads where assembled to create the virus genome, but virus reads were often present in multiple sequencing pools.

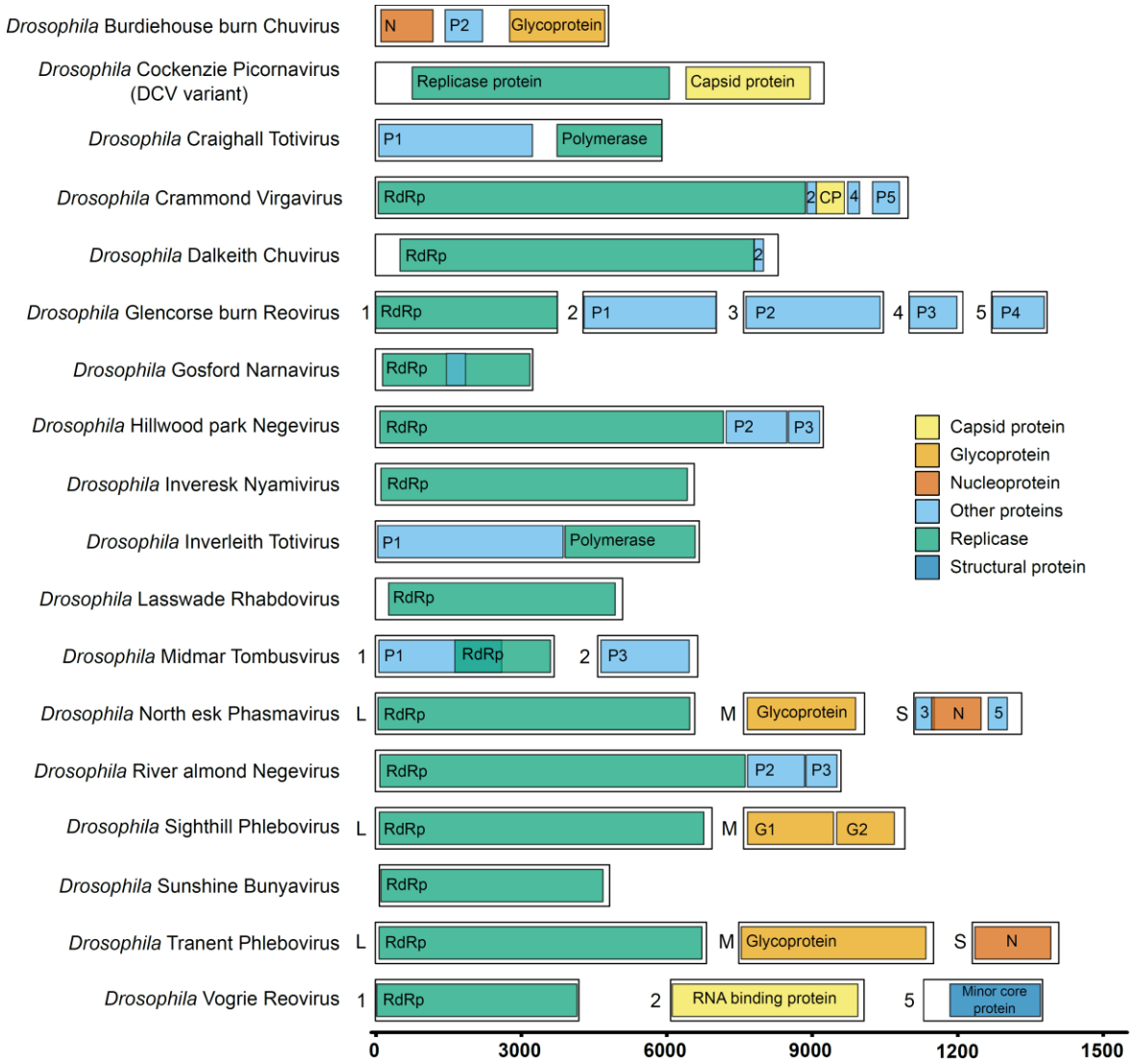


Fig. 2.3. Genome organisation of newly described viruses. Genome organisation of newly described RNA viruses, found in mixed species pools of *Drosophila.* All viral RNAs which exist as a negative sense strand are represented in their positive orientation. Colours indicate the hypothetical protein coding regions, and any functional annotations found by homology to other known virus proteins. See Fig. S2. **2** for an equivalent figure for the extended known virus genomes.

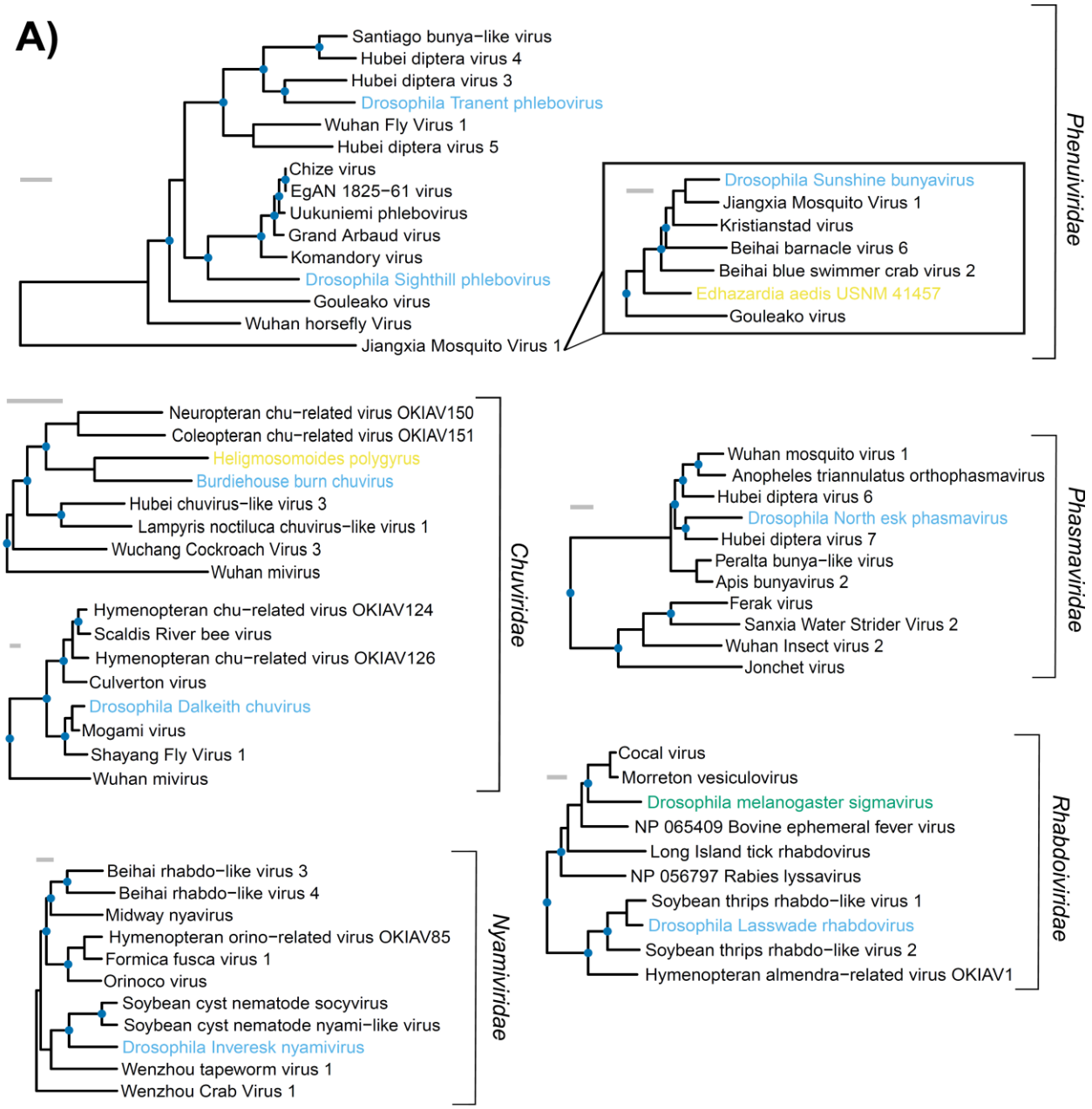
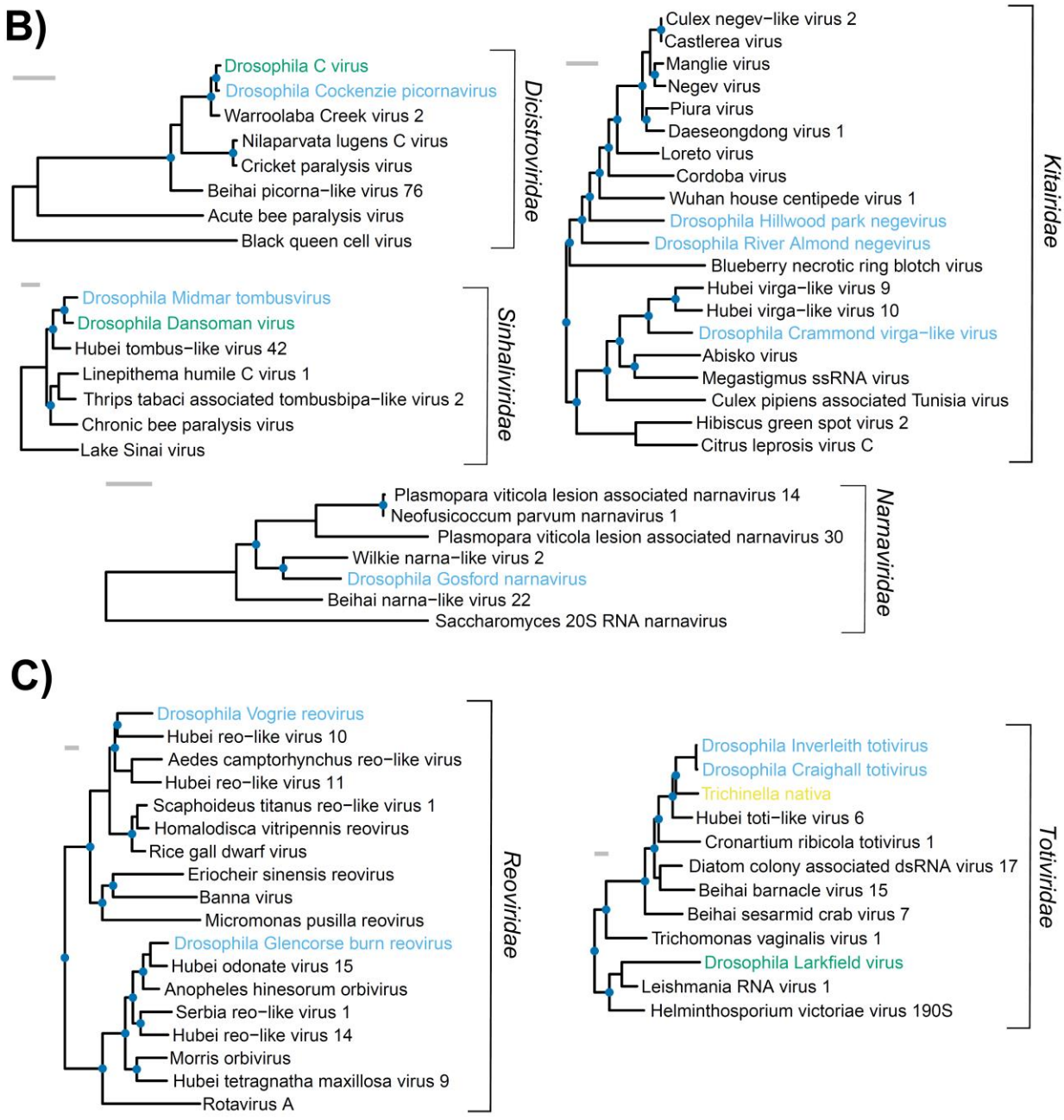


Fig. 2.4. (Continued on next page): Phylogenies of newly described Drosophila-infecting RNA viruses. Mid-point rooted phylogenies of viruses described in this chapter, containing the five closest BLASTp hits to the polymerase protein of each virus (or the glycoprotein for Burdiehouse burn virus), and representative viruses from each of the most closely related genera in the virus family. Families were allocated with reference to the 2019 ICTV report. The phylogenies are organised into A) Negative sense single-stranded viruses, B) positive sense single-stranded viruses and C) double-stranded viruses. The grey scale bars represent 0.5 amino acid substitutions per site. In each tree, viruses newly described in this chapter are labelled in blue, other viruses of Drosophila in green, and unannotated virus-like sequences in yellow. Node circles represent >= 60 bootstrap support. The inset tree indicates the phylogenetic relationships between viruses in separate genera within the Phenuiviridae.


B)
Drosophila C virus
Drosophila Cockenzie picornavirus
Warroolaba Creek virus 2
Nilaparvata lugens C virus
Cricket paralysis virus
Beihai picorna−like virus 76
Acute bee paralysis virus
Black queen cell virus
Dicistroviridae
Drosophila Midmar tombusvirus
Drosophila Dansoman virus
Hubei tombus−like virus 42
Linepithema humile C virus 1
Thrips tabaci associated tombusbipa−like virus 2
Chronic bee paralysis virus
Lake Sinai virus
Sinhaliviridae
Culex negev−like virus 2
Castlerea virus
Manglie virus
Negev virus
Piura virus
Daeseongdong virus 1
Loreto virus
Cordoba virus
Wuhan house centipede virus 1
Drosophila Hillwood park negevirus
Drosophila River Almond negevirus
Blueberry necrotic ring blotch virus
Hubei virga−like virus 9
Hubei virga−like virus 10
Drosophila Crammond virga−like virus
Abisko virus
Megastigmus ssRNA virus
Culex pipiens associated Tunisia virus
Hibiscus green spot virus 2
Citrus leprosis virus C
Kitaviridae
Plasmopara viticola lesion associated narnavirus 14
Neofusicoccum parvum narnavirus 1
Plasmopara viticola lesion associated narnavirus 30
Wilkie narna−like virus 2
Drosophila Gosford narnavirus
Beihai narna−like virus 22
Saccharomyces 20S RNA narnavirus
Narnaviridae
C)
Drosophila Vogrie reovirus
Hubei reo−like virus 10
Aedes camptorhynchus reo−like virus
Hubei reo−like virus 11
Scaphoideus titanus reo−like virus 1
Homalodisca vitripennis reovirus
Rice gall dwarf virus
Eriocheir sinensis reovirus
Banna virus
Micromonas pusilla reovirus
Drosophila Glencorse burn reovirus
Hubei odonate virus 15
Anopheles hinesorum orbivirus
Serbia reo−like virus 1
Hubei reo−like virus 14
Morris orbivirus
Hubei tetragnatha maxillosa virus 9
Rotavirus A
Reoviridae
Drosophila Inverleith totivirus
Drosophila Craighall totivirus
Trichinella nativa
Hubei toti−like virus 6
Cronartium ribicola totivirus 1
Diatom colony associated dsRNA virus 17
Beihai barnacle virus 15
Beihai sesarmid crab virus 7
Trichomonas vaginalis virus 1
Drosophila Larkfield virus
Leishmania RNA virus 1
Helminthosporium victoriae virus 190S
Totiviridae

936

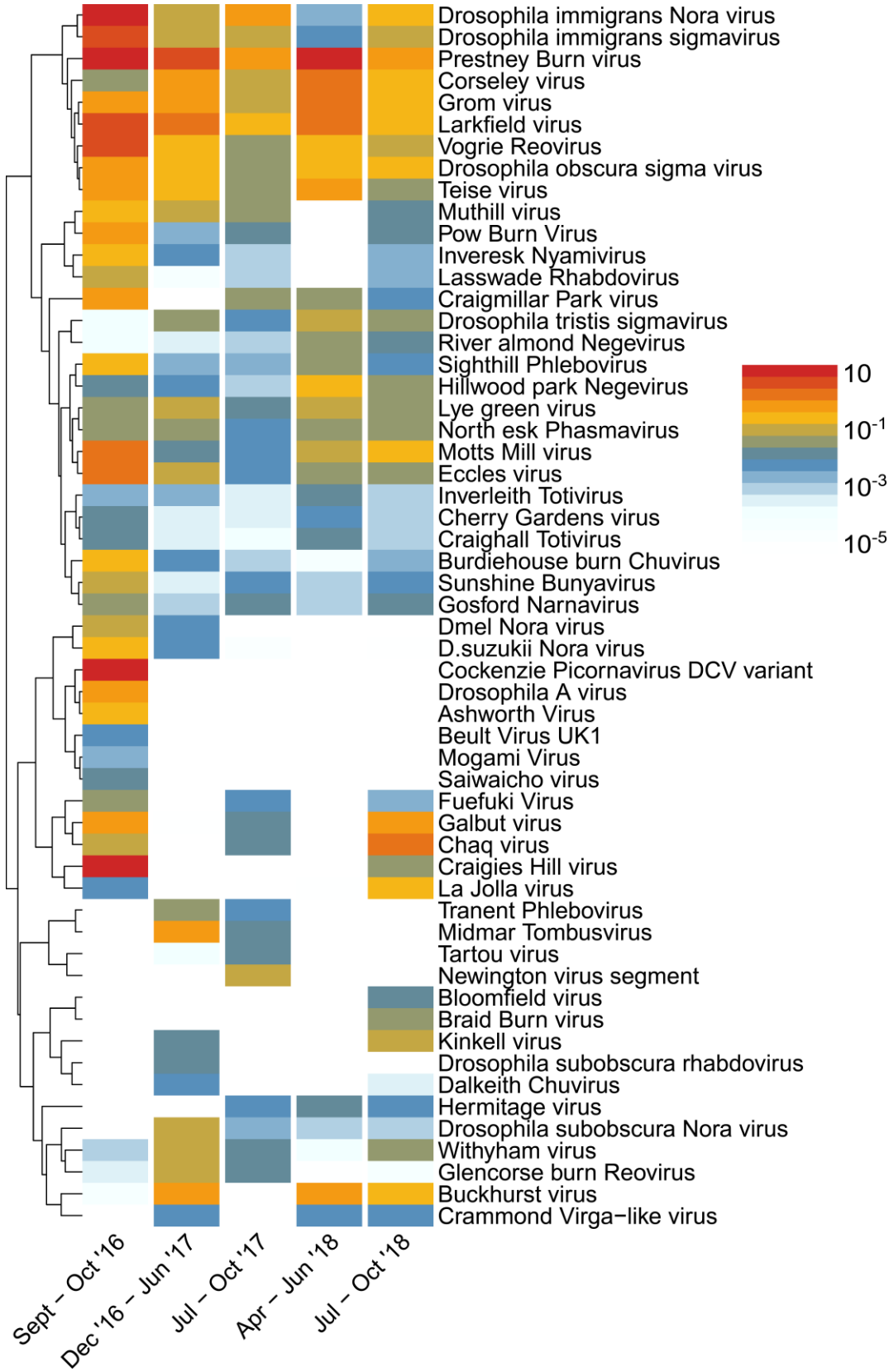


Fig. 2.5 Virus read numbers (relative to host Cytochrome oxidase 1, normalised for length). The plot shows the number of reads from each sequencing pool which map to newly described, and known, Drosophila viruses. Read numbers are normalised by target

sequence length and the number of reads mapping to the host COI gene, such that 1 is equivalent to equal numbers of virus copies to fly COI transcripts. See Fig. S2. **1** for read depth along each of the new virus genomes.

2.3.3 Host range of new and known *Drosophila* viruses

To determine which host species the new, and previously described viruses present in this system were infecting (i.e. their host range), I used previously published, and newly designed RT-PCR assays to screen species pools for 54 *Drosophila* viruses. For the new viruses, the co-occurrence of segments across species*sequencing pool combinations confirmed my assignment of segments to virus genomes (Fig. S2. ***3***). All but the Cockenzie Drosophila C virus variant (*D.subsilvestris*, September – October '16) were present in multiple sequencing datasets, and multiple host species (Fig. 2.6). Indeed, five viruses (Drosophila Midmar, Tranent, Burdiehouse burn, North esk and Craighall virus) infected five host species, the maximum across the dataset, four of these ranges encompassing both Drosophila subgenera. There was no obvious viral genome organisation, or viral family, with a propensity for large host ranges, as these five viruses encompassed four viral orders, and all three types of RNA virus genome organisation. Only six viruses displayed relative host specificity, Drosophila Crammond, Inveresk, Inverleith, Hillwood Park, River almond, and Sighthill viruses, which were restricted to the closely related Obscura group species.

Across new and known viruses, 41 viruses were infecting at least one host species in this system, 37 infected more than one species (90.2%), and 23 infected species from both subgenera of the genus *Drosophila* (Fig. 2.7). Common Drosophila laboratory infections, such as Drosophila X virus (Teninges *et al.* 1979), Drosophila A virus (Plus *et al.* 1975), and American Noda virus (Wu *et al.* 2010) were absent from all species and samples, corroborating their absence from Illumina reads. The presence of Drosophila C virus in the same single species and sequencing pool as the Cockenzie virus variant suggests that this is simply cross-priming to the variant, and that the commonly used laboratory strain (Jousset *et al.* 1977) is rare in the wild. Viruses previously identified as associated with *D. melanogaster*

and *D. simulans*, such as Drosophila melanogaster sigma virus (Fluriet 1982), Thika virus, and Kallithea virus (Webster *et al.* 2015), were also notably absent. Like the new viruses, very few of the previously described viruses showed high host specificity. Only La Jolla and Motts mill virus (Webster *et al.* 2015) were restricted to a single species (*D. melanogaster*), and only two viruses were restricted to the Obscura group (Tartou and Cherry Gardens virus). It's possible that the dominating host species in this community (*D. immigrans*) acts as a source for a significant proportion of the viral diversity, as all 19 *D. immigrans* infecting viruses also infected other species. For example, Drosophila immigrans Nora virus (DimmNV), and Drosophila immigrans sigma virus (DimmSV), both of which have only been previously described in *D. immigrans* (van Mierlo *et al.* 2014; Longdon *et al.* 2017), infect seven, and eleven species respectively.

There was no obvious pattern of association between viral genome types or orders and specific host species groups. Indeed, some closely related viruses show strikingly different breadths of host range. For example, Prestney burn virus, Grom virus, and Motts mill virus, described in Webster *et al.* (2015, 2016), are three closely related positive sense single-stranded RNA (+ssRNA) viruses (genome wide similarity 92% between Prestney burn and Grom, and 77-79% between Grom/Prestney burn and Motts mill). They infect eleven (Prestney burn), seven (Grom) and one species (Motts mill) respectively, suggesting that there isn't clade-wide level of transmissibility across these viruses, and that, whatever differentiates them, causes differences in their ability to infect novel host species.

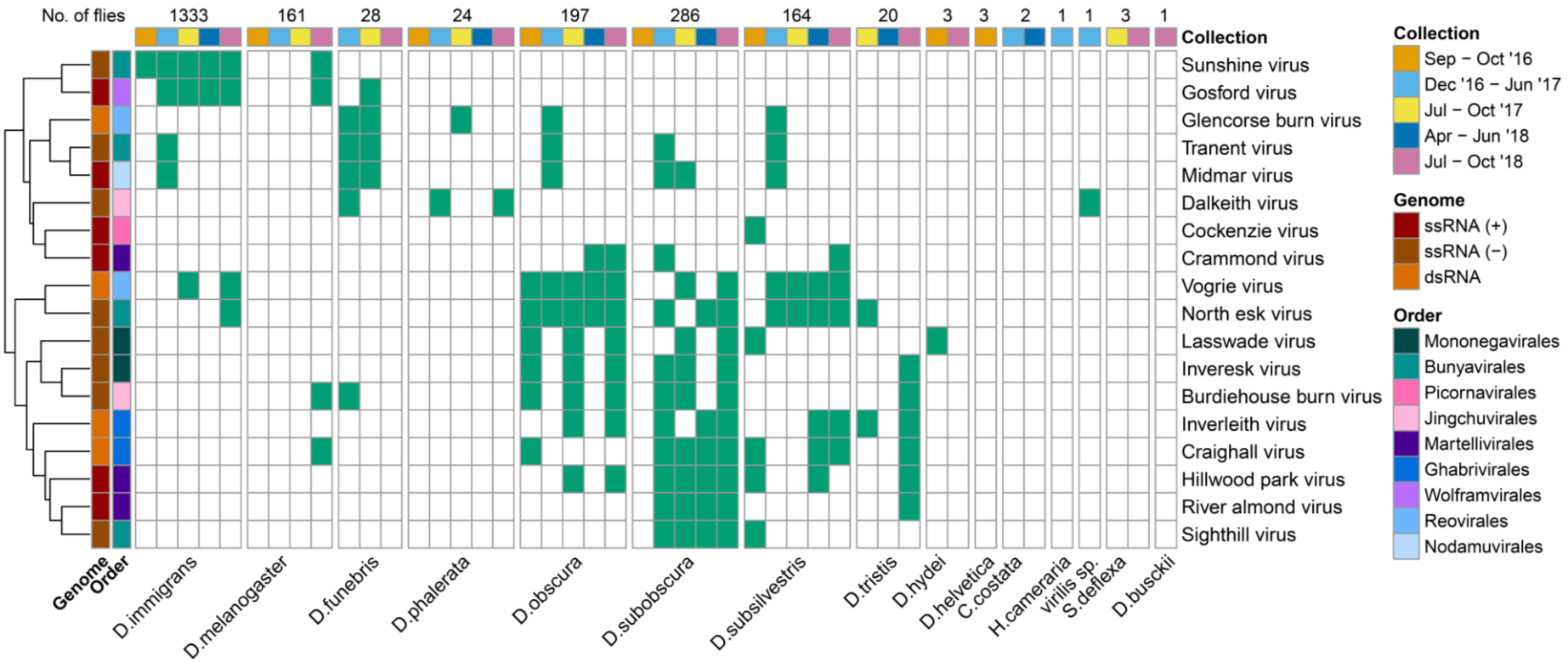


Fig. 2.6. Host range matrix for newly described Drosophila RNA viruses. The plot shows the presence/absence of each of the newly described RNA viruses from this chapter, as measured by RT-PCR on each of the sequenced pools of RNA. The number of each species collected across the five pools is denoted along the top of the matrix. The viruses were clustered by the similarity of their presence/absence data. A second version of this plot, which shows all of the viral segments, can be found in the appendix (Fig. S2. **3**)

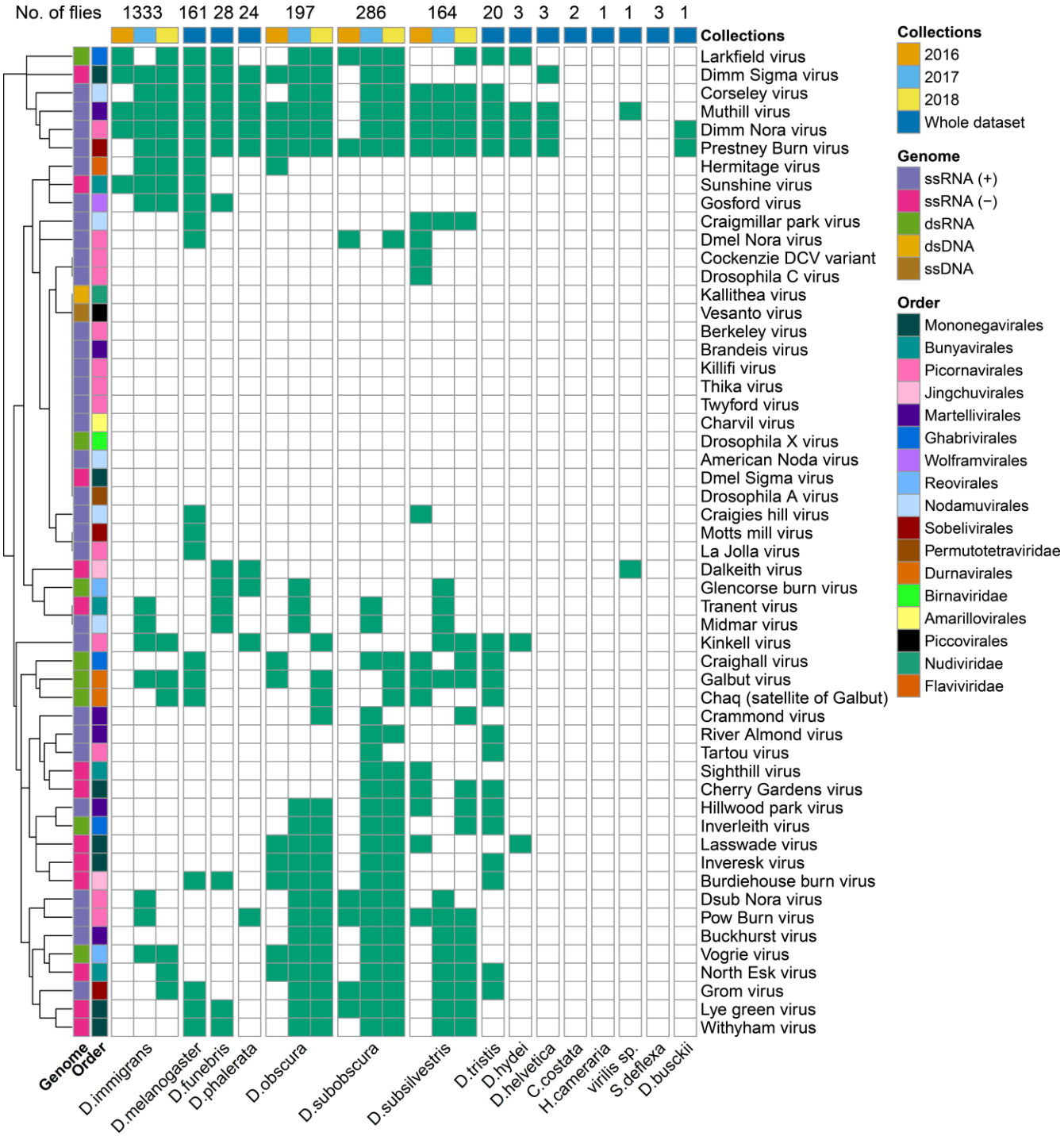


Fig. 2.7. The host range Drosophila viruses in my sampling area. The matrix shows the presence or absence of fifty four viral infections, assayed by RT-PCR, for each species collected. *Drosophila immigrans, obscura, subobscura* and *subsilvestris* were divided into years of collection for RT-PCR assays, due to high collection numbers. I selected viruses to assay for based on presence in the RNA sequencing pools from chapter 2, or their interest as lab contaminants or models. Viral order, and genome organisation is displayed as row annotations, taken from the 2019 ICTV report.

2.3.4 Prevalence of ten *Drosophila* viruses across species, season, and site

For ten viruses, I surveyed their prevalence (in addition to simple presence/absence) using PCR assays of individual flies and pools from every host species, site and month combination in the dataset. These viruses included two groups of three closely related viruses, to inspect their correlation in prevalence; Prestney burn, Motts mill and Grom viruses (see above), and DimmNV, Drosophila subobscura Nora virus (DsubNV) (van Mierlo *et al.* 2014), and Drosophila melanogaster Nora virus (DmelNV). They also included one new and one known virus with particularly wide host ranges; Muthill virus (Webster *et al.* 2016), and Tranent virus (described in this chapter), and, because of interest in their methods of transmission, DimmSV (see above), and Galbut virus (Webster *et al.* 2015), with its satellite – Chaq (Cross *et al.* 2020).

Overall viral prevalence (across all species, sites and months of sampling) varied between an estimated 18.4% (2 Log likelihood bounds = 16.6% - 20.3%) of flies infected by Prestney burn, and 0.226% (0.079% - 0.49%) for its close relative Motts Mill. The distribution of global prevalence was rather bimodal, with six of the ten viruses showing an estimated overall prevalence of less than 5%, and four (Prestney burn, Muthill, DimmSV, and DimmNV) estimated to infect >11% of flies. For all viruses apart from Motts mill virus, and Dsub Nora virus, prevalence varied significantly with host species, or host species and collection season. No viruses' prevalence varied by host species, collection season and collection site (Table 2.2). See Fig. 2.8 for the estimated prevalence of each virus across host species, and Fig. S2. ***5*** and Fig. S2. ***6*** for prevalence across months, and sites respectively. Like the host range data, the sets of closely related viruses showed extremely variable viral prevalence. Alongside the variation in the Prestney burn, Grom, Motts mill virus clade (see above), in the Nora viruses, overall virus prevalence ranged from 17.6% (2 LL = 15.5% - 19.7%) in DimmNV to 0.454% (2 LL = 0.224% - 0.803%) in DmelNV.

	Model1 (uniform prevalence)			Model2 (prevalence ~ species)			Model3 (prevalence ~ species + season)			Model4 (prevalence ~ species + season + site)		
Virus	Log Likelihood	No. of parameters	Akaike Weight	Log Likelihood	No. of parameters	Akaike Weight	Log Likelihood	No. of parameters	Akaike Weight	Log Likelihood	No. of parameters	Akaike Weight
Chaq satellite	-189.95	1	0.000	-129.44	15	0.319	-120.68	23	0.681	-41.96	170	0.000
DimmSV	-354.25	1	0.000	-258.02	15	0.904	-252.26	23	0.096	-162.66	170	0.000
DimmNV	-430.14	1	0.000	-402.03	15	0.001	-386.61	23	0.999	-257.54	170	0.000
DmelNV	-53.62	1	0.068	-37.01	15	0.931	-35.60	23	0.001	-21.52	170	0.000
Prestney Burn	-670.15	1	0.000	-387.70	15	0.449	-379.50	23	0.551	-264.53	170	0.000
Muthill	-389.14	1	0.000	-375.07	15	0.000	-358.56	23	1.000	-255.45	170	0.000
Tranent	-65.19	1	0.000	-49.78	15	0.002	-35.53	23	0.998	-17.90	170	0.000
Grom	-275.04	1	0.000	-177.49	15	0.970	-172.98	23	0.030	-102.47	170	0.000
Motts Mill	-28.66	1	0.593	-15.03	15	0.407	-15.00	23	0.000	-10.36	170	0.000
Galbut	-256.26	1	0.000	-191.46	15	0.006	-178.37	23	0.994	-93.94	170	0.000
DsubNV	-112.71	1	1.000	-108.84	15	0.000	-106.43	23	0.000	-64.03	170	0.000

Table 2.2. Log likelihood of viral prevalence variation under four models. The table shows the log likelihood (as calculated by a maximum likelihood estimation based on the presence/absence of a virus across different sized pools of flies) of viral prevalence if it was estimated separately or uniformly across three possible co-variates. The number of parameters shown were used to calculate the change in AIC (Akaike Information Criterion) between models, which takes into account the complexity of each model, and then the Akaike weights (relative likelihoods of the models) used to choose the model which maximised likelihood for each virus. The Akaike weights are coloured from red to green with the best fitting model for each virus coloured as green.

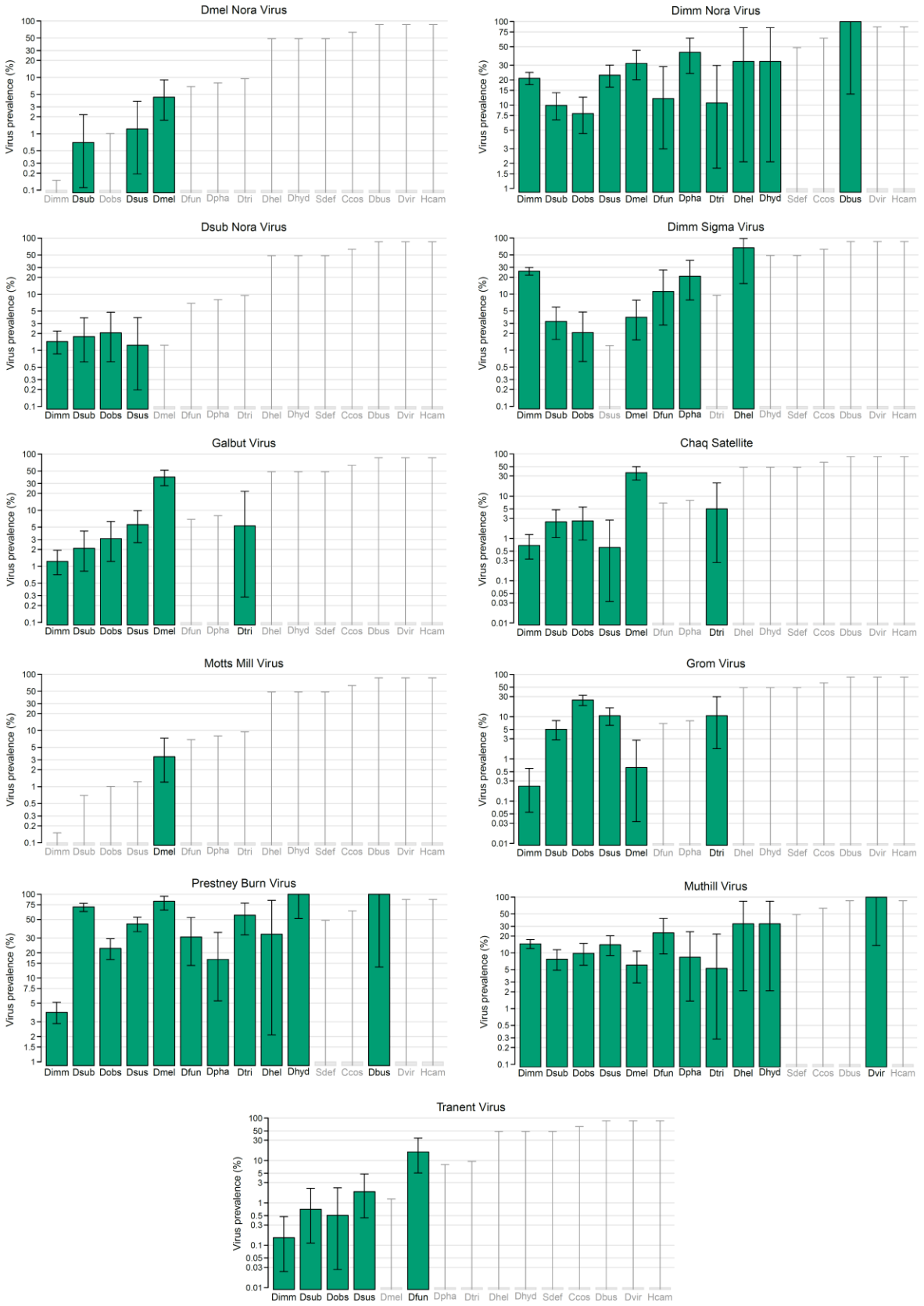


Fig. 2.8 Drosophila virus prevalence across species. The plots above show the estimated underlying prevalence of ten *Drosophila* viruses (and one *Drosophila* virus

satellite – Chaq) in the fifteen species collected. Greyed out bars indicate that prevalence was estimated to be < 0.01%. A log10 scale is used, that varies based on the lowest estimated lower bound for each species*virus combination. Error bars show 2 log likelihood bounds on the estimate. It should be noted that the number of flies collected varied considerably between species, as can be seen from the range covered by the error bars, and species are ordered by the numbers collected with the most common species on the left of each plot.

2.3.5 Genetic differentiation of *Drosophila* viruses across host species

Prestney burn virus, Motts mill virus, and Grom virus are three closely related +ssRNA viruses (Webster *et al.* 2015, 2016) that show markedly different host range, and prevalence, in the wild (see above). Their high sequence similarity leads to a danger that PCR primers could cross-prime to them, leading to us assigning the wrong virus to a host species. To check for cross-priming, and to see whether there was any significant divergence between viruses infecting different host species (for Grom and Prestney burn viruses) I sequenced regions of viral protein 1 (VP1) from each virus. I reassigned the only miss-identified sequences, which were 3 Prestney burn positives that were actually Grom virus. The rest of the PCR products were from the virus targeted, and were >98.6% identical to the published virus sequence (Motts mill - KP714076.1, Grom – KU754506.1, Prestney burn - KU754507.1) from Webster *et al.* (2015, 2016). Prestney burn virus (16 sequences) showed 11 SNPs (single nucleotide polymorphisms) in the 675bp region, 8 of which are non-synonymous, and Grom virus (30 sequences) showed 8 synonymous SNPs in the 804bp region. Both virus phylogenies were best inferred using an exponential population growth rate tree coalescent (exponential growth rate; Grom = 1.084, 95% HPD = 0.329 - 1.983, Prestney burn = 0.684, 95% HPD = 0.144 - 1.298) and showed panmixis of the viruses across six host species (Fig. S2. ***4***, A & B). This suggests that the previously published strains of these viruses are truly infecting multiple host species.

DimmNV has not previously been observed infecting species other than *Drosophila immigrans* (van Mierlo *et al.* 2014), and DimmSV is expected to have purely vertical transmission (Longdon *et al.* 2017). To see whether the observed infections in multiple species lie within the currently understood diversity of these viruses, and whether I was observing viral host shifts, or cross-priming to a closely related virus, I sequenced regions of the DimmNV, and DimmSV polymerase, and DimmSV capsid. All ten DimmNV nucleotide sequences were at least 99% identical to the previously published genome of DimmNV (KF24511.1) from van Mierlo *et al.* (2014) in this 640bp region. There was reasonable sequence divergence, with 14 SNPs among the sequences, all of which are non-synonymous. In the viral phylogeny, for which an exponential population size tree coalescent was supported (exponential growth rate = 1.272, 95% HPD interval = 0.1496-2.7422), there was no clear division of sequences by the six host species (Fig. S2. ***4*** C), suggesting an absence of genetic structure of the virus. All 12 DimmSV sequences, infecting *D.melanogaster* and *D.immigrans*, of both the L (630bp) and N genes (525bp) were >98.5% identical to a previously published sequence of DimmSV (KR822814.1) from Longdon *et al.* (2015b). There was little sequence divergence, with only 4 SNPs in the L gene region (1 non-synonymous), and 5 SNPs in the N gene region (4 non-synonymous). In the viral phylogeny (Fig. S2. ***4*** D), for which an exponential population size tree coalescent was supported (exponential growth rate = 1.711, 95% HPD interval = 0.1902-3.606), there was no clear division of sequences by host species, suggesting panmixis of the virus, and possibly horizontal transmission between host species.

2.3.6 Co-infections are not a rarity in wild multi-virus systems

Of the 416 individually assayed flies, 27.1% tested negative for the ten individual-based prevalence assay viruses, 40.6% were infected with one virus, and 32.3% were infected with multiple viruses (Fig. 2.9, A & B). The co-occurrence of the partitivirus Galbut virus (Webster et al. 2015), and it's satellite Chaq (Cross et al. 2020) only accounts for 14 co-occurrences (non-random co-occurrence, p=<0.0001*) among the 134 multiply infected individuals,

suggesting that co-infections are common in this system. I compared the observed frequencies of virus combinations with those expected based on a probabilistic model of species co-occurrence (Veech 2013) across all species (Fig. 2.9, C), and then within Dimm, Dsub, Dobs, Dsus and Dmel (the most commonly collected species). Many of the viral combinations that occur significantly more or less than expected across species are likely due to their reliance on the same, or different, host species. These are not likely to be good examples of competitive exclusion by the viruses (Vazeille et al. 2016), or viral facilitation *(eg. Kuwata et al. 2015)*. More interestingly, two viral combinations occurred significantly more often than expected within species. Muthill virus and DimmNV were significantly associated with one another in both Dsus and Dobs (Dsus – 7 co-infections, p=0.004*, Dobs – 4 co-infections, p=0.017*), and DimmNV and DimmSV were significantly associated in Dmel (4 co-infections, p=0.033*). However, sample size, which is low here because of the low numbers of multiply-infected, individually-assayed flies per species, would need to be larger to confirm if these viruses are facilitating infection by the other in some way. This analysis demonstrates that, in this system, multiple viral infections are not a rarity.

1074

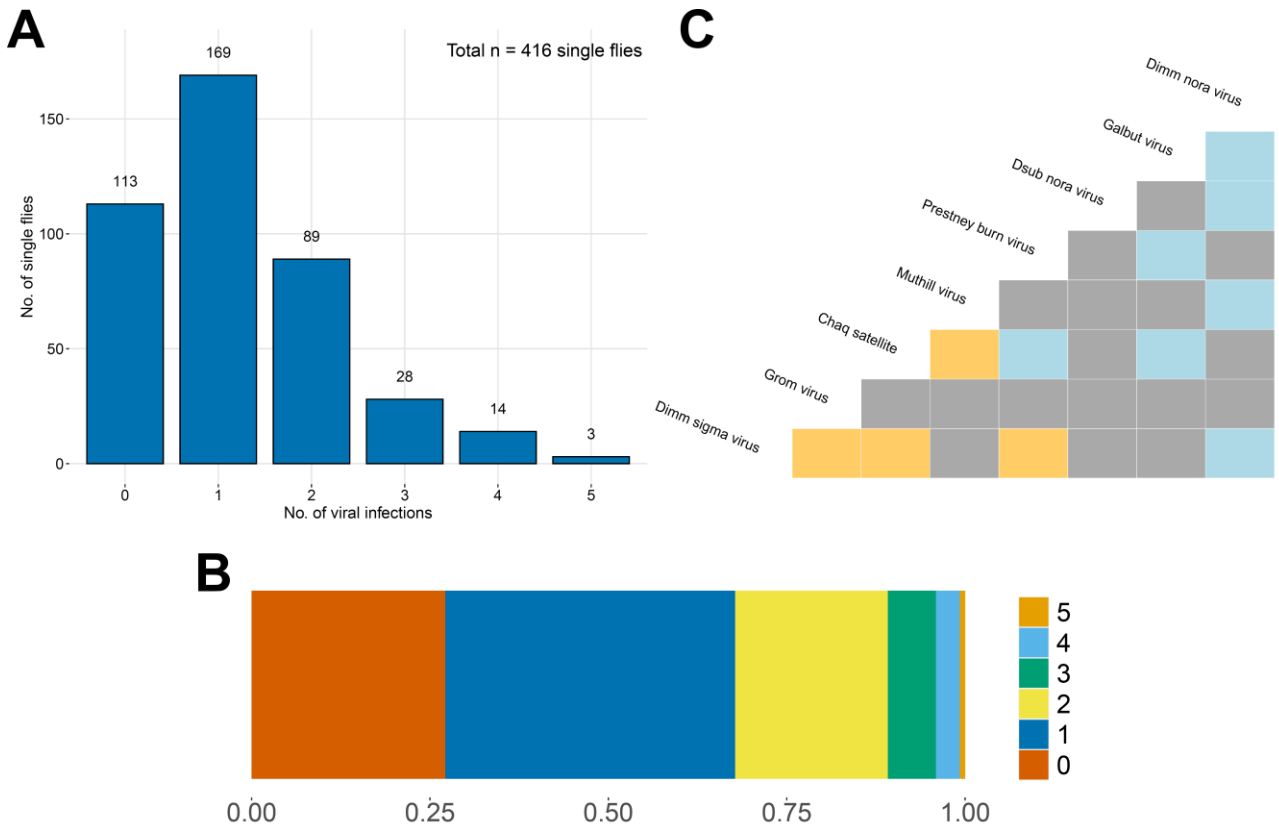


Fig. 2.9 The frequency of co-infection in wild Drosophila. The plots show the frequency and (A), proportion (B) of wild Drosophila infected with 0-5 viral infections. These figures were calculated from the 416 flies that were assayed for viruses individually. Panel C shows a co-occurrence matrix for the eight viruses which were common enough to analyse patterns of co-occurrence across all host species. Grey boxes indicate random patterns of co-infection, orange indicates significantly fewer co-infections than would be expected by chance, and blue indicates significantly more co-infections than would be expected by chance.

1075

2.4 Discussion

To understand and contextualise the evolutionary genetics of insect-virus interactions, we will need to answer several outstanding questions. Firstly, how consistent is the selective pressure imposed by viruses in the wild? This question can be thought of on two scales, one within populations, and one within species. Within populations, how does virus prevalence vary over time and space? And, within species, how constant are host-virus associations, and are most insect viruses generalists, or specialists? The second outstanding question is this, how are patterns of host shifting, and viral sharing shaped by these ecological dynamics? In this chapter, I use the *Drosophilidae*, and their naturally-occurring viruses to attempt to begin to answer some of these questions, and collate a dataset for future studies of the determinants of viral prevalence, and viral host range in insects.

2.4.1 Virus discovery and read abundance in Scottish drosophilids

I began by characterising previously unidentified viruses infecting Scottish drosophilids, identifying 17 new RNA viruses, and extending the genomes of 8 previously described viruses. The community of viruses infecting UK Drosophilids is already relatively well described, encompassing >70 *Drosophila*-associated viruses as a result of previous studies (eg. Longdon *et al.* 2010; Van Mierlo *et al.* 2014; Webster *et al.* 2015, 2016). However, by conducting an intensive surveys of this model system, and attempting to describe the full diversity of viruses infecting this sympatric species group, we can more quantitatively asses the burden of viral infection on insect hosts. In addition, the rapid pace of invertebrate metagenomics means that contigs not assignable as virus-like in previous metagenomic datasets might now be similar enough to newly released virus genomes to expand the diversity of *Drosophila*-associated viruses even further. For example, I was able to identify a *Nyami*-like –ssRNA virus (Drosophila Inveresk Nyamivirus) related to a recently described virus from Käfer *et al.* (2019), in which the phylogeny of the *Nyamiviridae* was significantly expanded. I also identified two viruses of the order *Jingchuvirales*, a clade described in Li *et al.* (2015). Of note, I screened a

similar panel of species to Webster *et al.* (2016), but unlike this study, did not use poly-A selection on reads, which biases discovery towards poly-adenylated genomes. This may be reason why I was able to identify four new Bunyaviruses (Table 2.1), which might have been selected out of previous surveys in this system because of their lack of poly-A tail. This could also be the reason why I identified no new *Picornavirales,* one of the virus families found to be particularly diverse and abundant in other metagenomic screens, in comparison to the two described in Webster *et al.* (2016).

Through the mapping of viral reads across the five sequencing datasets, and RT-PCR scans, I was able to assess not only which viruses were consistently present in this system, but also those that were consistently absent. There was a dearth of commonly used *Drosophila* viruses in the lab in both the RT-PCR survey, and metagenomic survey, illustrating their rarity in this temperate system. Some of this can be attributed to the rarity of Dmel in my collections, but it perhaps reinforces the importance of studying native viruses in *Drosophila* in co-evolutionary experiments which use this panel of species, as the currently used lab viruses do not represent a shared co-evolutionary history with these hosts. However, I did identify a variant of Drosophila C virus (Drosophila Cockenzie DCV variant – see Table 2.1) infecting a single pool of *Drosophila subsilvestris* in September-October 2016. DCV has been used in studies of insect immune defence (Ferreira *et al.* 2014), viral clearance (Mondotte *et al.* 2018), and predictors of host shifts (Longdon *et al.* 2015a; Imrie *et al.* 2021), and causes eventual mortality through intestinal obstruction (Arnold *et al.* 2013b). For experimental studies of *Drosophila*-virus interactions, it's perhaps reassuring that a variant of this commonly used virus is present in the wild, but its rarity is likely to reflect the rarity of such highly virulent viruses in the wild, and perhaps, that infections of *Drosophila* might represent spillover from another reservoir hosts. DNA viruses were also seemingly absent from this system, consistent with their rarity in previous *Drosophila* metagenomic datasets (Webster *et al.* 2015).

The benefit of repeatedly sequencing a single viral community is that it allows the identification of viruses which are each other's closest relatives, and therefore could be used for future

studies of viral divergence, co-divergence with their hosts, and host switching (eg. Longdon *et al.* 2011). For example, I identified two viruses of the *Totiviridae* family (Drosophila Craighall, and Drosophila Inverleith Totivirus), with 82.2% genome-wide similarity. I designed primers specifically to differentiate between these two viruses in RT-PCR surveys, and found them to infect a similar panel of host species (Dsus, Dobs and Dsil, though Inverleith virus also infects Dmel). Repeated sequencing of viruses like this could reveal historic patterns of divergence on their viral genomes, and host shifts. Two other previously identified groups of closely related *Drosophila* viruses which could be used for these kinds of studies are the Nora virus cluster (DimmNV, DmelNV and DsubNV - +ssRNA Picornavirales), and a cluster of closely related +ssRNA viruses related to Poleroviruses and Sobemoviruses (Grom, Prestney burn virus, and Motts Mill virus). The Nora viruses have already been used in the lab to examine the determinants of host range, via the host-specific action of viral suppressors of RNAi (van Mierlo *et al.* 2014), but no previous studies have compared and contrasted their host range and prevalence in the wild.

2.4.2 Drosophila virus host range

To characterise the host range of not only these closely related groups of viruses, but the wider *Drosophila* virosphere, I screened the 15 Drosophila species collected, over three years, for 54 viruses using RT-PCR. By doing this, I aimed to understand whether most viruses in this system are generalists, or specialists. The results of this screen reflects the broad propensity of RNA viruses to infect multiple, sympatric, but distantly related host species in an insect community. Of the 41 viruses present in this system, 37 infected more than one species, and 23 infected species from both subgenera of the *Drosophilidae*. Indeed, single-host, single-virus associations are an extreme rarity. I was not able to examine the prevalence of the majority of these viruses across species, and therefore the likelihood that some of these infections represent spillover, or 'apparently multi-host pathogens' (Fenton & Pedersen 2005). However, this generalism contrasts with the high level of host specificity seen in a survey of mosquito viruses across species within a national park in the Côte d'Ivoire, where ~80% of

viruses infected a single host species (Hermanns *et al.* 2021). In contrast to this, in this system only two viruses, La Jolla and Motts mill virus (Webster *et al.* 2015), were restricted to a single species (*D. melanogaster*), and only two viruses were restricted to the Obscura group (Tartou and Cherry Gardens virus).

Of note, I observed infections in multiple species for two viruses previously thought to be highly host specific, DimmNV, and DimmSV. DimmNV was thought to be highly host specific due to the host-specific action of its viral suppressor of RNAi (van Mierlo *et al.* 2014). However, in this system DimmNV showed particularly wide host range, at an estimated prevalence of >5% in eleven species (Fig 2.8) across both subgenera of the *Drosophilidae*, without host specific divergence in its polymerase protein sequence (see Fig S2.4). 200 DimmNV reads were found in a sequencing of *D.suzukii* in Medd *et al.* (2018), which could be attributed to barcode switching (Ballenghien *et al.* 2017), however it's also possible that this virus has indeed expanded its host range. The three Nora viruses also showed extremely variable host range, and prevalence, despite their close relatedness, making them prime candidates for future studies of viral divergence and evolution in multi-host systems. DimmSV, which was previously thought to be obligately vertically transmitted, also displayed a wide host range, with viral RNA detectable by PCR in seven species. The phylogeny of sigma viruses shows frequent host shifting (Longdon *et al.* 2011c), typical of *Rhabdoviridae* (Geoghegan *et al.* 2017) and so it's possible that this virus has expanded its range, or is now being transmitted horizontally.

2.4.3 Variation in Drosophila virus prevalence by host species, and season

One of the key outstanding questions in the study of insect-virus co-evolutionary dynamics is how consistent host-virus associations are, both in evolutionary time (ie. Host range and host shifting), but also over ecological timescales. I found virus prevalence to vary significantly not only across host species (for eight viruses), but also by season for six viruses. This is consistent with other studies which find variation in insect virus prevalence with site (Odindo 1982), and environmental factors (Myer & Johnston 2019), which could in turn drive host species abundance. The variation in prevalence I observed between host species could be

structured by host phylogeny, and in particular, the shared co-evolutionary history of host species driving common immune defences (Longdon *et al.* 2011a; Faria *et al.* 2013). It's unlikely that viral phylogeny is driving some of these prevalence patterns, based on the wide variation in prevalence seen between closely related viruses (eg. Prestney Burn and Motts Mill viruses). However, more explicit co-phylogenetic mixed models (eg. Hadfield *et al.* 2014) would be needed to evaluate the power of host, and virus phylogenies, to drive both patterns of host range, and prevalence in this system.

2.4.4 Ecological drivers of virus dynamics

By characterising the ecological dynamics of multi-host, multi-virus systems, it allows us to think about the ways in which this might drive evolutionary dynamics. For example, the significant variation in virus prevalence observed for six viruses between the early and late sampling seasons, and between host species (Table 2.2), suggests that perhaps the different community compositions of these seasons drive variation in virus prevalence, and therefore virus-driven selection. Changes in host community composition, influenced by habitat disturbance, were found to drive changes in viral prevalence in a study of mosquitoes and their native viruses (Hermanns *et al.* 2021), and perhaps similar eco-evolutionary dynamics could be at play here. *D.subobscura* overwinters as an adult, but has no reproductive diapause, unlike the *Drosophila* sub-genus species *D.phalerata* and *D.funebris* (Lumme & Laakovaara 1983), which is perhaps why members of the Obscura group dominate the community in the early collecting season. This changing community composition could potentially mean that in temperate systems like this one, viruses might need to be more generalist to persist in the system year round.

The over-representation of Drosophila immigrans in these collections (59.85% of flies collected were Dimm) also lead me to hypothesise that *D. immigrans* acts as a source for a significant proportion of the viral diversity, as all 19 *D. immigrans* infecting viruses also infected other species. However, sampling density is likely to bias this assumption, as I have, most likely, far more completely characterised the diversity of viral infections in Dimm in comparison

to other species. A virus at 30% prevalence in Dimm would infect ~400 flies in this dataset, but only ~8 in *D.funebris,* demonstrating the difficulty in assessing the source-sink dynamics possibly at play in this system. In order to rigorously evaluate the determinants of viral prevalence and host range in this system, we will need to use models which control for this potential sampling bias, whilst evaluating the strength of environmental and phylogenetic predictors.

2.5 Conclusions

In this chapter, I aimed to demonstrate that by examining the ecological dynamics of multi-host, multi-virus systems, we can contextualise the results of studies of insect-virus co-evolutionary dynamics. I used monthly sampling of the *Drosophilidae,* and their naturally-occurring viruses to collect a multi-year dataset of virus host range, and prevalence in a wild insect community. Few other studies have previously attempted to examine *Drosophila* virus prevalence outside of Dmel, and none have characterised it's variation in space, and time. I used this dataset to identify 17 new *Drosophila*-associated viruses, some of which are key candidates for studies of co-divergence, and co-evolution. I also evaluated the host range of 41 *Drosophila*-infecting viruses and found that >90% of them infect multiple species. This suggests that focussing studies of insect-virus co-evolution on single-host, single-virus interactions will now allow us to accurately characterise the dynamics of the insect virosphere. Additionally, I find evidence that *Drosophila* viruses vary significantly in their prevalence between host species, and between seasons. This suggests that not only evolutionary, but also ecological factors could drive inconsistency in virus-driven selection pressure on insect genomes.

3 Prevalence variation and genetic diversity in DNA viruses infecting European *Drosophila*

Work in this thesis chapter has been accepted for publication in the journal Virus Evolution as part of Wallace et al. (2020), and the current manuscript can be found at https://doi.org/10.1101/2020.10.16.342956. Virus discovery work in this manuscript, and sequence analysis up to and including the competitive mapping against contaminants in bowtie2 was done by Darren J. Obbard, and therefore forms part of the introduction to this chapter, the further sections of which focus on the downstream analysis of prevalence, and genetic diversity in DNA viruses done by MW. Sample collections and DNA sequencing runs were organised and paid for by the DrosEU consortium. The R-INLA models for the spatial GLMM analyses were built with invaluable help from Gregory F. Albery, who also wrote and advised on use of the ggregplot package in R to analyse their output.

3.1 Introduction

The use of naturally-occurring host-virus combinations allows the study of 'typical' wild co-evolutionary dynamics between insects and their viruses, which might help us to understand the evolution of insect-vectored viruses with economic and public health impacts. However, using the *Drosophilidae* and their viruses as a model system for insect-virus co-evolution requires that these models represent the dynamics of the whole virosphere, which not only includes RNA viruses, but also DNA viruses.

Before 2020, very few DNA viruses of *Drosophila* had been identified. However, metagenomic sequencing has now built the current number of known *Drosophila*-infecting DNA viruses to sixteen. In 2011, Unkless (2011) identified the first native DNA virus of the *Drosophilidae*, *Drosophila innubila* Nudivirus (DiNV), a large dsDNA virus that causes reduced offspring production in both *D. innubila* and *D. falleni*. This was followed by the discovery of two more

DNA viruses of *Drosophila,* through sequencing of multi-species pools. The first, Kallithea virus (KV), is a large ~153kb dsDNA Nudivirus of *D. melanogaster* (Webster *et al.* 2015), and the second, Invertebrate iridescent virus 31, in *D. obscura* and *D. immigrans*, (Webster *et al.* 2016 - reads almost identical to *Armadillidium vilgare* iridescent virus from Piégu *et al.* 2014). Isolation, and characterisation of the fitness effects of Kallithea virus, which is likely transmitted faecal-orally, found it to reach a high titre on infection in both sexes of *D.melanogaster*, reducing survival in males, and movement and late-life fecundity in females (Palmer *et al.* 2018b).

In 2020, two studies together identified an additional 13 *Drosophila*-associated DNA viruses from the collection and sequencing of 167 pools of male Dmel (some contaminated with *D. simulans*) from >30 locations across Europe in 2014-2016, as part of a concerted effort to characterise continent-wide population structure, and selection in the hosts (Kapun *et al.* 2020; Wallace *et al.* 2020). This included the description of 5.4kb Linvill Road virus, a virus likely to be a member of the ssDNA *Parvoviridae*, and most closely related to the unclassified *Haemotobia irritans* densovirus (Ribeiro *et al.* 2019). Linvill Road virus lies within the 'Densoviruses', a sub-family of *Parvoviridae* which are thought to infect a diverse range of insect hosts (Porter *et al.* 2019). Densoviruses are specifically associated with rapidly dividing cell types, and can be highly pathogenic (Tijssen *et al.* 2016). Through expanded sampling and long-read sequencing, (Wallace *et al.* 2020) also extended the genome of the putative segmented ssDNA Bidnavirus, Vesanto virus (first described in Kapun *et al.* 2020), identifying 12 putative segments all between 3.3 and 5.8 kb in length. Overall, the DNA viruses of *Drosophila* now represent a wide variety of viral genome organisations, and give us the opportunity to investigate the way that these different types of insect infecting viruses, evolve, vary in prevalence and are structured within populations.

DNA viruses represent a middle ground between RNA viruses (small genomes, high mutation rates) and eukaryotes (larger genomes, low mutation rates) and can be more akin to bacteria or archaea in terms of their evolution. Large, double stranded DNA (dsDNA) viruses usually

show much slower mutation rates than RNA viruses, on the order of $1x10^{-7}$-$1x10^{-8}$ mutations per site per replication (Duffy 2018). This is caused by the higher fidelity of their DNA polymerase, and their larger genome size compared to most RNA viruses. In comparison, single stranded DNA (ssDNA) viruses (eg. Shackelton *et al.* 2005; reviewed in Duffy *et al.* 2008), show mutation rates only slightly lower than RNA viruses at 10^{-5} to 10^{-6} mutations per site per replication (Duffy 2018), possibly due to the lack of a complementary DNA strand (Tijssen *et al.* 2016).

The higher mutation rate of ssDNA viruses in comparison to dsDNA viruses results in higher levels of ssDNA virus genetic diversity within and among hosts (eg. Ge *et al.* 2007) in comparison to dsDNA viruses. This is because neutral genetic diversity, measured as mean pairwise differences at synonymous sites ($\prod_s$) is expected to depend upon the mutation rate (μ). Neutral genetic diversity also depends on the rate of drift, measured as its reciprocal, effective population size - N_e. In viruses, N_e is not only reduced when virus prevalence is low, but also by the bottlenecks induced by invasion of a new host (reducing within-host N_e). This means that if DNA virus prevalence is low in *Drosophila,* this, and their lower mutation rates, might lead to lower genetic diversity in their DNA viruses of *Drosophila* in comparison to RNA viruses.

Indeed, in comparison to the high prevalence of some *Drosophila*-infecting RNA viruses, the DNA viruses of *Drosophila* appear to be much less prevalent. Some RNA viruses of *Drosophila* are able to exist at an extremely high prevalence in the wild, such as Galbut virus, an endogenous copy of this virus was detected segregating in European populations of Dmel (Wallace *et al.* 2020). Though Drosophila innubila Nudivirus (DiNV), a DNA virus, shows a similar prevalence to Galbut virus in North America (36% across all males and 25% in females - Unkless 2011), all DNA viruses discovered in European collections of *Drosophila* have an estimated prevalence of <5% in the wild, and most are found in only a single sample, equating to an estimated prevalence of 0.015% (Webster *et al.* 2015; Wallace *et al.* 2020). In American populations of *Drosophila*, DiNV exhibits recurrent evolution of a high titre variant (Hill &

Unckless 2020), but the low prevalence of European *Drosophila* DNA viruses might restrict their adaptive potential, constraining such dynamics. The restraint that low prevalence effects on DNA virus evolution in European Drosophila would depend on the consistency of this low prevalence in space, and time. However, studies of the predictors of fine and broad-scale spatiotemporal variation in wild insect virus prevalence, for both RNA and DNA viruses, are rare.

The Hytrosaviruses are a notable exception to this. In the Salivary gland hypertrophy (SGH) viruses (dsDNA) of tsetse flies and other Diptera, prevalence is low in the wild (0.2-5.4%), similar to the DNA viruses of *Drosophila*, but can vary by location, trap site, and even seasonally within trap sites (see review Lietze *et al.* 2011). This perhaps indicates that we should expect very fine-scale variation in low-prevalence insect DNA viruses. In RNA viruses, studies on the drivers of virus prevalence have focused on viruses infecting pollinators (such as honeybees, reviewed in Manley *et al.* 2015; McMahon *et al.* 2018), or viruses vectored by insects to agricultural animals and humans. Studies like these can access the predictors of fine-scale variation in viral presence / absence, and characterise and account for spatial and temporal autocorrelation in their datasets using spatially structured random effects (Blangiardo *et al.* 2013). For example, Myer & Johnston (2019) analysed the local scale predictors of West Nile virus (WNV) incidence in traps (pools per trap) of mosquitoes over 15 years, finding spatial patterns of WNV incidence, alongside increased likelihood of incidence in wetlands, areas with low vegetation and high urban development. We are unaware of any studies that have used this framework to characterise spatial and temporal variation in the incidence of insect DNA viruses.

In this study, I aimed to utilise the data produced from the extensive large-scale sampling and genome sequencing employed by the DrosEU consortium (in Kapun *et al.* 2020; Wallace *et al.* 2020), to characterise both ecological and evolutionary dynamics in insect DNA viruses. I characterised the spatial and temporal dynamics of DNA virus incidence for the three most common DNA viruses of European *Drosophila* described in (Kapun *et al.* 2020; Wallace *et al.*

2020): the dsDNA Nudivirus Kallithea virus, and two ssDNA viruses, Linvill Road virus and Vesanto virus. I was also able to identify spatial and seasonal hotspots in the incidence of a recently integrated endogenous copy of Galbut virus in the Dmel host genome. Although the pooled nature of this dataset limited the use of some traditional population genetic approaches, I also describe the patterns of genetic diversity in these viruses infecting Dmel and *D. simulans* (Dsim).

3.2 Methods

3.2.1 Sample collection, sequencing and prior analyses

A dataset was analysed which included 167 pooled samples of 33-40 male *Drosophila melanogaster*, collected across 47 different collection sites in Europe (latitude = 35.07 to 62.55, longitude = -8.41 to 38.72) by members of the DrosEU consortium. In total, 6668 flies were collected from natural or semi-natural habitats using fruit and yeast baited traps between June 2014 and November 2016, and identified to species by morphology. Of the 47 collection locations, flies were collected from 30 in more than one year, and from 19 in all three years. Flies were also collected from several sites in both the early and late *Drosophila* breeding season. Flies from each collection were pooled, before being transported and stored in ethanol at -20ºC or -80ºC. DNA extraction from each of the pools was performed using a phenol-chloroform based protocol. DNA was sequenced using 151nt paired-end Illumina reads in three blocks; Block 1 (2014 samples) on the Illumina NextSeq platform, and Block 2 and 3 (most 2015 samples, and remaining 2015 samples and 2016 samples) on the HiSeq X platform. Illumina reads from each of the samples were trimmed (Martin 2011; Krueger 2015) and competitively mapped to the genomes of common *Drosophila* contaminants, *Drosophila* host species, and *Drosophila* DNA viruses using bowtie2 (Langmead & Salzberg 2012), retaining only the best mapping position for each read. This mapping allowed the identification of putative virus contigs for virus discovery, and the identification of samples infected with *Wolbachia* endosymbionts, and samples contaminated with species other than Dmel. Thirty pools were contaminated with a non-melanogaster *Drosophila* species, 30 with *D.simulans* (Dsim), one with *D. phalerata*, one with *D. testacea.*

Fourteen DNA viruses of *Drosophila* were found in at least one pool in this dataset, (quantified by the viral copy number relative to fly genome copy number). For more specific details on the field collection strategy, DNA extraction and sequencing of the data analysed in this study, see supplementary material of Kapun *et al.* (2020), and for specific collection information for

all samples from 2014-2016 analysed in this study, and initial mapping of reads, see Wallace *et al.* (2020). The final genetic diversity datasets presented in this chapter consisted of reads competitively mapped to all viruses present in the dataset. The final viral prevalence datasets included per-sample relative viral copy number for all viruses present in the dataset, along with data on their collection location, and levels of contamination (data at https://doi.org/10.6084/m9.figshare.14161250.v1).

3.2.2 Spatial and temporal variation in viral copy number

The relative viral copy number for each virus (or EVE): sample combination was converted to binary presence / absence data, using a threshold of 1% to indicate infection in a pool (or 0.1% for the Galbut EVE). For viruses that infected more than one sample, I then characterised patterns of spatial and temporal heterogeneity in viral site-level incidence by fitting generalised linear mixed models (GLMMs) with binomially distributed response variables in R v4.0.2 (R Core Team 2020) using the linear modelling package R-INLA (Blangiardo et al. 2013). Using R-INLA allowed me to control for and quantify spatial autocorrelation using the applied Integrated Nested Laplace Approximation (INLA), using a Stochastic Partial Differentiation Equation (SPDE) approach. INLA SPDE is a deterministic Bayesian approach which evaluates the spatial covariance between two sites remaining after accounting for all other model parameters (more details in Lindgren & Rue 2015). I used this package to add a spatially distributed random effect into the models if its inclusion was supported by a change in 2 Deviance information criterion (DIC), indicating an improvement in model fit. I was able to build models of incidence variation for the three most common DNA viruses in this dataset: Kallithea virus (93 infections), Linvill Road virus (21 infections), and Vesanto virus (114 infections), and the Galbut EVE (42 samples with the endogenous copy). I had insufficient data to build reasonable models for the remaining two multi-sample viruses in the dataset, Esparto virus and Viltain virus (both infected only 5/167 pools). From the binary data, I also

estimated global fly-level prevalence using a maximum likelihood algorithm, based on a model of pooled Bernoulli trials (eg. Speybroeck et al. 2012). This function searches across a range of prevalence values to identify the prevalence which maximises the likelihood of the observed number of pools/single flies testing positive, given the sizes and number of pools.

GLMMs for virus incidence included four explanatory variables: relative number of reads mapping to Wolbachia (continuous), the level of *D. simulans* contamination, measured as the percentage of reads mapping to *D. simulans* Ago2 (continuous), sampling season (categorical with two levels: early and late season) and year (categorical with three levels: 2014, 2015 and 2016). I also included collection location as a random effect to account for any sampler, or site effects in the data. The final model formula was;

Viral prevalence ~ Wolbachia + Year + Season + %Dsim + f(sampling location)

ΔDIC supported the inclusion of a spatially distributed random effect to account for spatial autocorrelation in the models for Kallithea virus, Linvill Road virus, and Galbut EVE incidence. For Kallithea and Linvill Road viruses, a separate spatial mesh was not supported for each year or season, indicating consistent spatial hotspots of infection (see Fig. 3.1, A & B for plots of the spatial fields used). In contrast, ΔDIC supported the inclusion of a separate, uncorrelated, spatial field to account for variation in incidence of the Galbut EVE in the early and late sampling seasons (see Fig. 3.1 C & D).

I evaluated models (looking at ΔDIC, effect sizes, and the range of decay in spatial autocorrelation) and plotted spatial fields using functions from the ggregplot (https://github.com/gfalbery/ggregplot), and ggplot2 (Wickham 2009) packages in R. Specifically, I checked the robustness of fixed effects to spatiotemporal autocorrelation by plotting the consequence of the inclusion of a spatial or spatiotemporal (separate spatial meshes per season) random effect on the fixed effect posterior estimates. To examine the distance over which pool infection status is correlated, I plotted the speed of decay of spatial autocorrelation in space (kappa/range) for each virus or EVE where its inclusion was

supported in the model. Finally, I plotted the best-fit (based on ΔDIC) spatial fields, converting the projections for site-level prevalence (*q*) to individual fly-level prevalence (p) as p = 1 – exp (log(1-q)/ n). Where *n* is the number of flies in the pool (assumed to be 40 across the study).

3.2.3 Genetic diversity

For my analysis of genetic diversity, I used SAMtools v1.10 (Li *et al.* 2009) to exclude virus samples with less than 10-fold coverage across 95% of their genome, and viruses with less than 6 samples. I then filtered each of the competitively mapped bam files for properly paired reads, mapping to the remaining, and three most common, DNA viruses (Kallithea, Linvill Road and Vesanto virus). Following this, I quality trimmed read-ends with cutadapt 2.10 (Martin 2011), using a minimum read length of 75bp, and a quality threshold of 18, accounting for the two-colour chemistry reactions in samples from Block 1 which were sequenced on the NextSeq platform. The trimmed reads were then remapped to the virus reference genomes using bwa mem (Li 2013), which performs local alignment of reads and will store multiple primary alignments for a query sequence. For Vesanto virus, I remapped to multiple haplotypes of some of the segments, as they were sufficiently diverse. To reduce cross-mapping, I also removed the terminal inverted repeat regions (TIRs) from the ends of the Vesanto segment contigs prior to remapping.

Next, I filtered alignments to include only primary alignments, and those with a Phred-scaled mapping quality (MAPQ) of greater than 30. PCR and optical duplicates were then identified and removed using picard v.2.22.8 MarkDuplicates (Broad Institute, 2020) to reduce the effect of any amplification of reads during library preparation and sequencing. Finally, I assessed coverage using bedtools v2.26.0 (Quinlan & Hall 2010). Sample bam files with a per-position read depth of less than 25 across 95% of the non-TIR virus genome were excluded from further analyses. After identifying the most common haplotype for each Vesanto virus segment in each of the samples, I filtered the competitively mapped bam files for all haplotypes of the segment and then remapped the Vesanto segment reads (local alignment) such that only a

single variant of each segment was represented in each sample. I repeated all pre-processing steps (cutadapt trimming, duplicate removal, and read filtering) on these new single-haplotype alignments.

To calculate global genetic diversity for each virus, I created single 'whole population' alignments by merging all the sample alignments for each virus or segment haplotype. To speed up computational processing, I ensured that no single sample contributed more than an average 500-fold coverage of reads, and to even coverage across the viral genomes, I also down sampled the bam files to their median read depth using SAMtools (Li *et al.* 2009). As misalignment around indels can cause miscalled bases, the single sample and whole population bam files were re-aligned around indellic regions using GATK v3.8 (Van der Auwera *et al.* 2013). To calculate allele frequencies at each position of the virus genomes for diversity analyses, I generated allelic counts from the single sample and whole population alignments using SAMtools (Li *et al.* 2009). I used a relatively conservative minimum base quality of 40 and minimum MAPQ of 30, to limit the influence of sequencing error on the allele counts, and down-sampled the single samples to a maximum read depth of 500. I then identified and masked (with a window size of 5bp) indellic positions supported by at least 7 reads using popoolation (Kofler *et al.* 2011a). I generated allelic counts for variant positions in each of the whole population alignments using popoolation2 (Kofler *et al.* 2011b), limiting the search to single nucleotide polymorphisms (SNPs) with a minor allele frequency of >=1%, again to limit the effect of sequencing error.

To calculate nucleotide diversity at synonymous (π_S) and non-synonymous (π_A) sites I first calculated the average number of synonymous and non-synonymous sites in each gene in the virus genomes. To do this, I used the SNPs for each virus to calculate a transition-transversion ratio in R v4.0.2, and then, the average number of non-synonymous sites per codon (Kofler *et al.* 2011a). For each virus genome, I then used the average number of non-synonymous sites per codon, and the gene sequences, to calculate the average number of synonymous and non-synonymous sites in each gene. I identified synonymous and non-

synonymous SNPs, and generated allelic counts, in both the whole population and single samples using popoolation (Kofler *et al.* 2011a) with a minimum coverage of 5 for any variant positions. I then filtered out any SNPs with a minor allele frequency of less than 1% and calculated mean π_A and π_S as the average number of pairwise differences between sequences across synonymous and non-synonymous sites. I also calculated π_A and π_S for each of the genes, and across intergenic regions.

I plotted neutral site diversity (π_S and intergenic π) across the Kallithea virus genome, using 2 and 5kb sliding windows. The sliding windows were calculated using adapted code from the R package 'evobiR' (Anderson et al. 2019) so that the sliding 'mean' = $\sum$ (per site intergenic and synonymous site wise nucleotide diversity)/ $\sum$ (average synonymous + intergenic sites in the reference genome). I identified genes which significantly deviated from the expected ratio of non-synonymous to synonymous SNPs using a binomial test, with the expected proportion of non-synonymous SNPs defined by the proportion of non-synonymous sites (ie. does the ratio deviate from that expected with a dN/dS of 1, neutrality). I only considered genes with at least 5 SNPs in this analysis, and corrected p-values for the rate of false discovery using a Benjimini Hochberg correction (Benjamini & Hochberg 1995).

I recorded any single samples with no SNPs (minimum allele frequency 3%) to inspect viral haplotypes, and recombination. However, there were fewer than four Linvill Road and Kallithea virus samples with no SNPs, meaning there were insufficient invariant samples to analyse recombination. For Vesanto virus, only some segment haplotypes showed >4 invariant samples, and as this wasn't consistent across the segmented virus genome, I was also unable to analyse recombination for this virus.

3.2.4 Structural diversity in the Kallithea virus genome

To catalogue short InDel polymorphisms, and in particular their division across coding and intergenic regions, I identified indellic positions in each of the infected samples, using popoolation2 (Kofler et al. 2011b). I found InDels to be so rare across the Vesanto virus

segments and Linvill Road virus, as to provide insufficient power for any kind of analysis. For Kallithea virus, I designated positions as coding or non-coding according to the reference genome (KX130344.1) in R v4.0.2 (R Core Team, 2020). I then tested for any association between the coding status of a position and the chance of an InDel ever being detected there (ie. a gap is supported by at least 5 reads in at least 1 sample), using a chi-square test for independence. I plotted variation in the abundance of InDels across the Kallithea virus genome as percentage of samples with support for a gap at each position. These percentages were calculated along a sliding window (window size = 10, step size = 1) using R package evobiR (Anderson et al. 2019).

R and shell scripts used in the genetic diversity, InDel, and spatiotemporal incidence analyses, and to create plots from this data, can be found on the Figshare repository for Wallace *et al.* 2020 (https://doi.org/10.6084/m9.figshare.14161250.v1).

3.3 Results

3.3.1 DNA virus prevalence varies in space and time

I estimated global fly-level prevalence to be low for all three viruses analysed in this study. The most common was Vesanto virus at 2.83% (2.31-3.42%), the next most common Kallithea virus at 2.01% (1.62-2.47%), followed by Linvill Road virus, at 0.34% (0.21-0.5%). These estimates assume that the underlying virus presence / absence observations are independent of one another. In reality, viral prevalence may be correlated in time and space, or dependent on other ecological predictors. I examined this non-independence using GLMMs, which examined the effect of contamination by D.simulans, the amount of Wolbachia, year of collection and collection season on viral incidence. I found contrasting levels of spatiotemporal predictability, and dependence on the fixed effect predictors for each virus. In all models, fixed effects estimates were only slightly modified by the addition of a spatial, or spatiotemporal random effect. No estimates were reduced in their significance, indicating that these effects are robust to spatial autocorrelation (See Fig. 3.2 for effect sizes under non-spatial, spatial and spatiotemporal models). I found insignificant local-scale predictability in viral incidence (location random effect accounted for < 0.02% of variance in all final models).

Kallithea virus, and Linvill Road virus site-level prevalence is correlated in space, but not in time. These two GLMMs were improved by the addition of a spatial random effect (ΔDIC from non-spatial models: Kallithea = -13.59, Linvill Road = -17.2). However, neither model was improved by allowing the spatial field to vary between year, or season, indicating consistent hotspots of infection. The spatially distributed random effects in the Kallithea and Linvill Road virus models accounted for 19.17% and 33.21% of variance respectively, and final model DIC for Kallithea and Linvill Road virus was 220.68 and 101.83 respectively. In Fig. 3.1 panel A and B, I show the spatial fields of the SPDE random effects for these two viruses. Though the distribution of these two viruses is quite different, and Kallithea virus was found to be more common in the study area, they showed a similar, long geographic distance over which site-

level prevalence was spatially correlated (See Fig. 3.3 for autocorrelation decay across space). Spatial autocorrelation halved ('halving distance') after a distance of 7.58 and 8.59 coordinate units, and had a range of 23.1 (6.79-63.85) and 26.9 (7.13-75.23) degrees for Kallithea and Linvill Road virus respectively. In contrast to the other two virus models, my model of Vesanto virus was not improved by the addition of a spatial random effect. Given this insufficient support for a spatial model, the fact that that I had no *a priori* expectation that spatial distribution would vary between years, and the reduction in power it would induce in each category, I did not attempt to fit separate spatial random effects per year for Vesanto virus.

Vesanto virus was the only virus with prevalence that varied significantly over time, with a higher prevalence in 2015 and 2016 in comparison to 2014 (change in log odds effect size for 2015 = 1.274, 0.423 - 2.161, for 2016 = 1.426, 0.5 - 2.401). See Fig. 3.2 for all effect size estimates. The final model DIC for Vesanto virus was 208.03. Linvill Road virus also showed a significantly higher prevalence in samples with a higher percentage of reads mapping to D.simulans (change in log odds = 8.5, 2.91 - 15.44). For all viruses, neither collection season nor the amount of Wolbachia explained a significant proportion of variation in site-level prevalence.

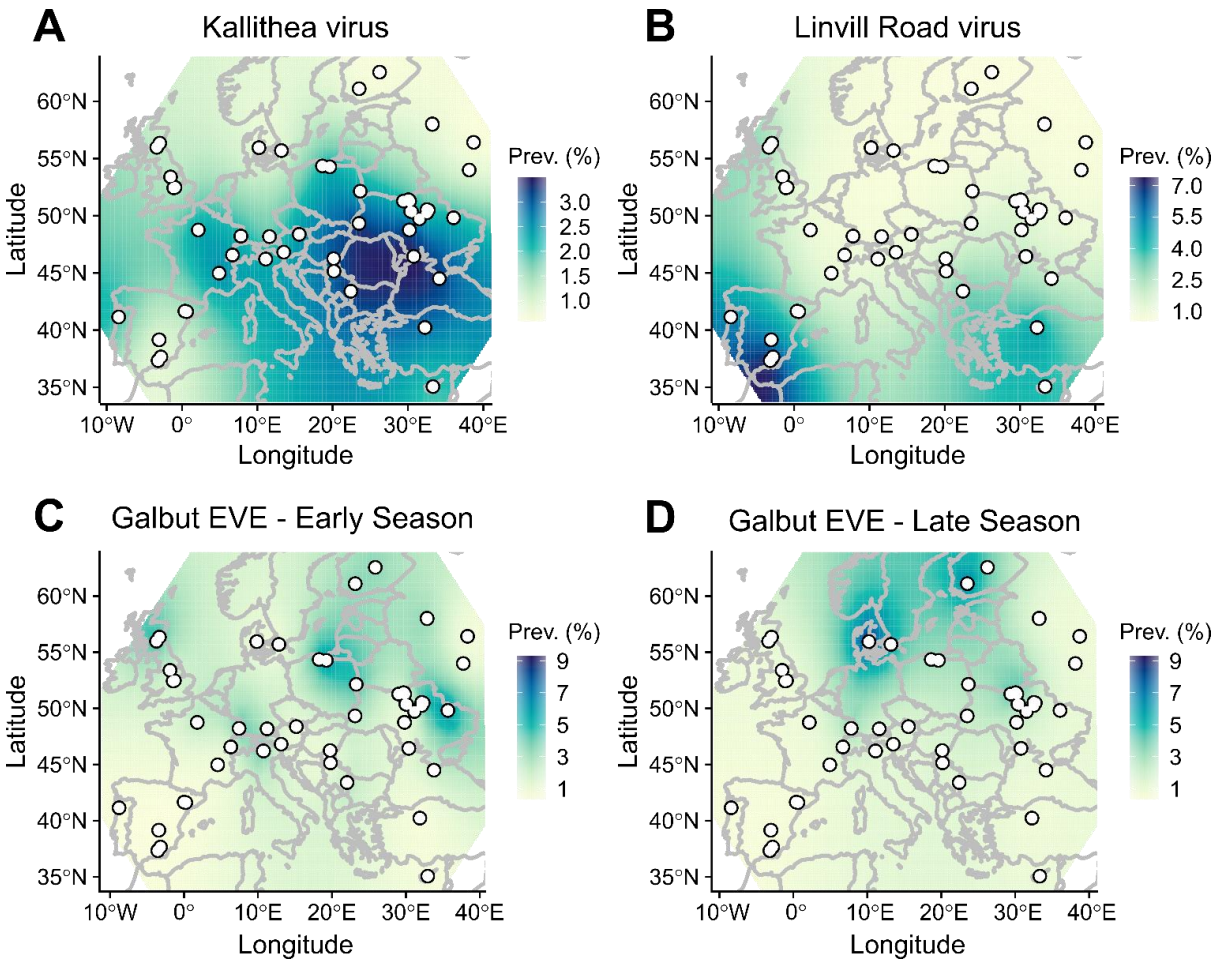
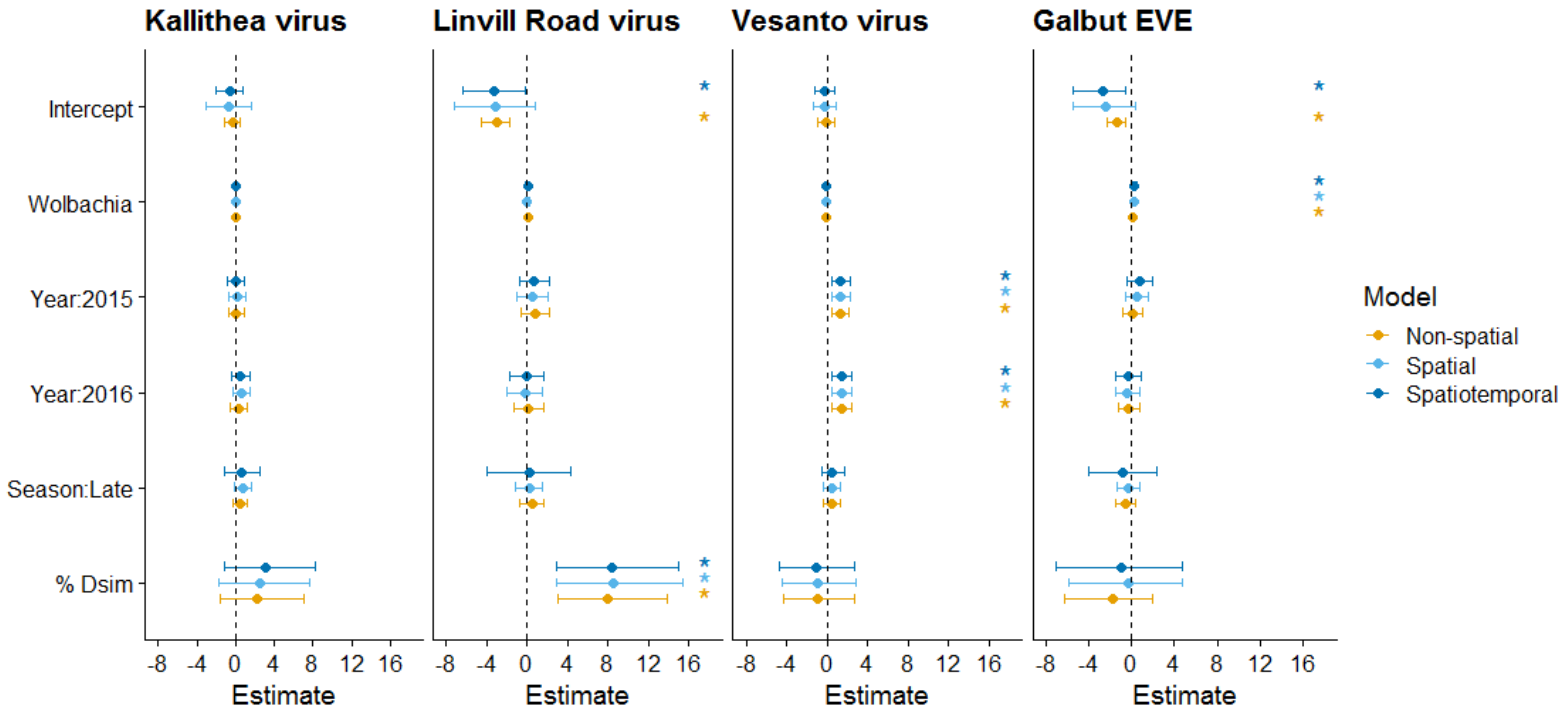


Fig. 3.1 Spatial variation in the copy number of Kallithea (A) and Linvill Road virus (B), and the Galbut EVE (C & D). Maps of Europe on coordinate scales depicting sampling sites, and projections of the spatially distributed SPDE random effects (spatial fields) for the models of virus copy number which were improved by their inclusion. Legend indicates the colour scale for plotting, which shows the predicted prevalence of the virus or EVE at the single fly level (Prev. = prevalence). The model of copy number variation for the Galbut EVE was improved with the addition of a separate spatial field for early and late season collections, the projections for which are shown in panel C & D.



1565 **Fig. 3.2**. **Comparison of the fixed effects estimates from each of the INLA models for viral prevalence.** The plot shows a comparison between the posterior estimates from the base model set (Non-spatial), the base model set + a spatial random effect (Spatial), and the base model set + spatial random effect which varied between seasons (Spatiotemporal). Points show the mean effect estimate (as a change in log odds) and the error bars show the 95% credible interval. Estimates for categorical variables represent the change from the missing category of each variable (Year: 2014 and Season: Early). Asterix denote fixed effects which were deemed significant (95% credible intervals did not overlap 0).

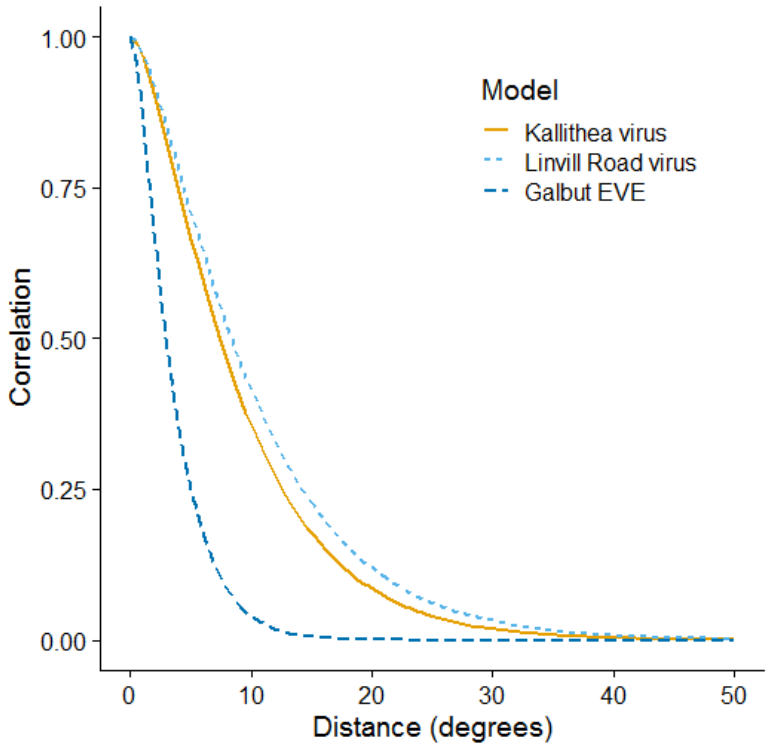


Fig. 3.3 **The range of spatial autocorrelation acting on Kallithea virus, Linvill Road virus and the Galbut EVE.** The plot shows the decay in spatial autocorrelation over coordinate units for the three INLA models in which the inclusion of a spatially distributed random effect was supported by ΔDIC. The steeper the curve of the line, the faster autocorrelation decays, and the shorter the distance over which viral prevalence is correlated across space.

3.3.2 A recently integrated endogenous copy of Galbut virus is spatially and seasonally distributed

In 42 of the 167 samples, based on a threshold of 0.1% relative viral copy number, an endogenous copy of Galbut virus was found in the host genome. The site-level copy number of this EVE varied predictably both in space, and in time, with a different spatial distribution in the early and late sampling seasons. The GLMM of incidence variation for the Galbut EVE was improved by the addition of a separate, uncorrelated spatially distributed random effect

per collection season (Final model DIC = 167.42, ΔDIC from non-spatial = -25.51, ΔDIC from spatial = -5.52). The spatiotemporal component of this model accounted for a large proportion (45.97%) of variance. The EVE was most commonly in Dmel sampled around Scotland, Poland and Ukraine in the early sampling season, and more northerly, around Sweden, Denmark and Finland in the late sampling season. (See Fig. 3.1 panel C and D). The spatial distribution of the Galbut EVE appears patchier than that of Kallithea or Linvill Road virus, reflected in the shorter range (See Fig. 3.3) of spatial autocorrelation (10.20 degrees, 2.27-30.17, halving distance 3.03 degrees). *Wolbachia* contamination marginally increased the likelihood of finding the Galbut EVE (change in log odds = 0.231, 0.074 - 0.410) (Fig. 3.2).

3.3.3 Genetic diversity

I examined nucleotide diversity in the three of the most common DNA viruses of *Drosophila*; Kallithea, Linvill Road and Vesanto virus. I identified 923 single nucleotide polymorphisms (SNPs) across the total Kallithea virus population and a sum of 15,132 unique SNPs across the 44 infected samples. In Kallithea virus the vast majority of variants are globally low in frequency. 13,291 SNPs were private to a single sample (ie. singletons), and 1,100 were private to two samples (doubletons). Indeed, over 90% of Kallithea virus local SNPs were found in only 1 or 2 samples. Across the 44 infected samples, I found 28 positions in the Kallithea virus genome to be polymorphic in at least 15 samples, potentially maintained through balancing selection, or possibly in the middle of a selective sweep. The majority of these SNPs were in intergenic regions, were synonymous, or were in genes encoding proteins with unknown functions. For example, in intergenic regions, there were 12 segregating polymorphisms upstream of the first gene in the Kallithea virus assembly (KX130344.1), a putative protease (AQN78547.1). There is also a segregating site upstream of a vlf-1-like protein encoding gene (AQN78580), which contains an integrase domain sometimes important for recombination and repair, and a synonymous SNP segregating in a gene encoding a vp91-like protein, which contains a chitin binding domain (ACH96236.1). Only two segregating SNPs showed non-synonymous effects, a Serine to Proline mutation in 17

populations in position 60123 of the DNA Helicase 2-like protein (AQN78620.1), which unwinds DNA and is critical in its repair, and a Proline to Threonine mutation in 15 populations in position 40776 of the GrBNV gp37-like protein (AQN78629).

Virus	Total π_s	Total intergenic π	Total π_A/π_s	Mean local π_s	Mean local intergenic π	Mean local π_A/π_s
Kallithea virus	0.15%	0.14%	0.39	0.04%	0.05%	0.58
Linvill Road virus	1.45%	0.84%	0.10	0.21%	0.14%	0.19
Vesanto virus (mean across segments)	1.16%	0.47%	0.20	0.28%	0.13%	0.24

Table 3.1 Genetic diversity of three DNA viruses infecting European Drosophila. The table shows calculated values for the average number of nucleotide differences per site (π) between sequences in the 'Total' virus population and local (within sample) virus populations. π_A, π_S and intergenic π values were calculated in R v 4.0.2 as $\sum$ (per site nucleotide diversity at synonymous, intergenic or non-synonymous sites) / average number of synonymous, intergenic or non-synonymous sites. Total and mean local π_A/π_S values were calculated as mean (π_A) / mean (π_S). All Vesanto virus summary statistics were calculated as a mean across all segments and segment haplotypes.

The global population of Kallithea virus showed relatively low nucleotide diversity at neutral, or nearly-neutral sites (π_s = 0.15%, and intergenic π = 0.14%) in comparison to that of its host (Kapun *et al.* 2020), with little pattern to their variation (Fig. 3.1, panel A). Within single population pools, neutral site diversity was ~3-fold lower still (Table 3.1), equating to high population differentiation (F_{st} = 0.73). At both the local and total population levels, the average level of constraint on evolution was low; total π_A/π_s = 0.39, and mean local π_A/π_s = 0.58. In the total population pool, 14/95 genes (which had at least 5 SNPs) showed a π_A/π_s >1 (Table S3. 1), indicating diversion from neutrality. These 14 genes include those encoding a lef-3-like protein (AQN78596.1), a trypsin-like serine protease-like protein (AQN78562.1), an odv-e56-like protein (AQN78623.1), a GrBNV gp81-like protein, an Ac92-like protein (AQN78621.1)

and a DNA helicase-2-like protein (AQN78620.1). No genes showed statistically significant evidence of positive selection (eg. high π_A/π_s, and dN/dS). However, 9 genes showed evidence of negative selection, i.e. a lower ratio of non-synonymous to synonymous SNPs than would be expected with neutrality, including genes encoding a lef-9-like (AQN78613.1), and a lef-8-like (AQN78600.1) protein and a gp83-like protein (AQN78594.1). It's likely that the evolution of more of the Kallithea virus genes is constrained, but I had insufficient power to detect it.

I also characterised indel variation in the Kallithea virus genome by identifying short indels within each sample. Across 44 infected samples, I identified 2289 positions in the Kallithea virus genome as indellic. However, consistent with the global rarity of Kallithea virus SNPs, most of these were rare. Only 195 indels had reached an appreciable frequency in the global population (were found in at least 50% of the samples). Most indellic positions (1774) were found in intergenic regions of the genome, as expected due to selection against mostly deleterious coding indels (Fig. 3.4, panel B). A chi-square test for independence on the number of indels found in cds and intergenic regions found a strong positive association between intergenic regions and finding InDels (X-squared = 3236, df = 1, p-value < 2.2e-16).

In Vesanto virus, an ssDNA segmented Bidna-like virus, I identified 4,059 SNPs across all segments and divergent segment haplotypes in the total population, and a sum of 5,491 distinct SNPs across all infected samples. Again, the vast majority of SNPs were low in frequency. 4,235 were private to a single population, and 721 were private to two populations. I found both global and local diversity at synonymous sites to be higher in Vesanto virus than Kallithea virus, by a factor of ~7, with total π_s and mean local π_s calculated at 1.16% and 0.28% respectively (Table 3.1). Again, these statistics also show a within population diversity which is much lower than total population diversity, equating to a high F_{st} of 0.76. I identified a total of 441 indellic positions across all Vesanto segments and segment haplotypes, and a mean of only 13.78 indels per segment haplotype (3.3-5.8 kb) (range = 0 – 63 indels). The majority of these indels were rare, with 339 occurring in only 1 or 2 samples.

In Linvill Road virus, an ssDNA Densovirus which is the rarest of the three viruses analysed, I identified 178 SNPs across the total virus population. Across the 13 samples with reasonable coverage, I identified a sum of 253 distinct SNPs, again a large majority of which were private to one or two populations (209 singletons, and 30 doubletons). In comparison to Kallithea virus, Linvill Road virus displayed higher levels of nucleotide diversity at synonymous sites in both the global and local populations (Total πs = 1.45%, Mean local πs = 0.21%) (Table 3.1). Local neutral site diversity was ~7-fold lower than total diversity, though intergenic π was less similar to π_s at both a local and global level (Total intergenic π = 0.84%, and Mean local intergenic π = 0.14%). This again suggests that most SNPs are rare, or recent, and that high levels of population differentiation exist (F_{st} = 0.86). The levels of constraint on evolution in both the total and local populations of Linvill Road virus were intermediate, with total and local π_A/π_s at 0.10 and 0.19 respectively (Table 3.1), and mean total π_A/π_s per gene at 0.10 also. No genes in the Linvill road genome showed a total π_A/π_s greater than 1 (Table S3. 1), most likely due to the small, compact genome size of Linvill Road. I identified only 3 indellic positions in the samples infected with Linvill Road, all of which were present only in a single sample.

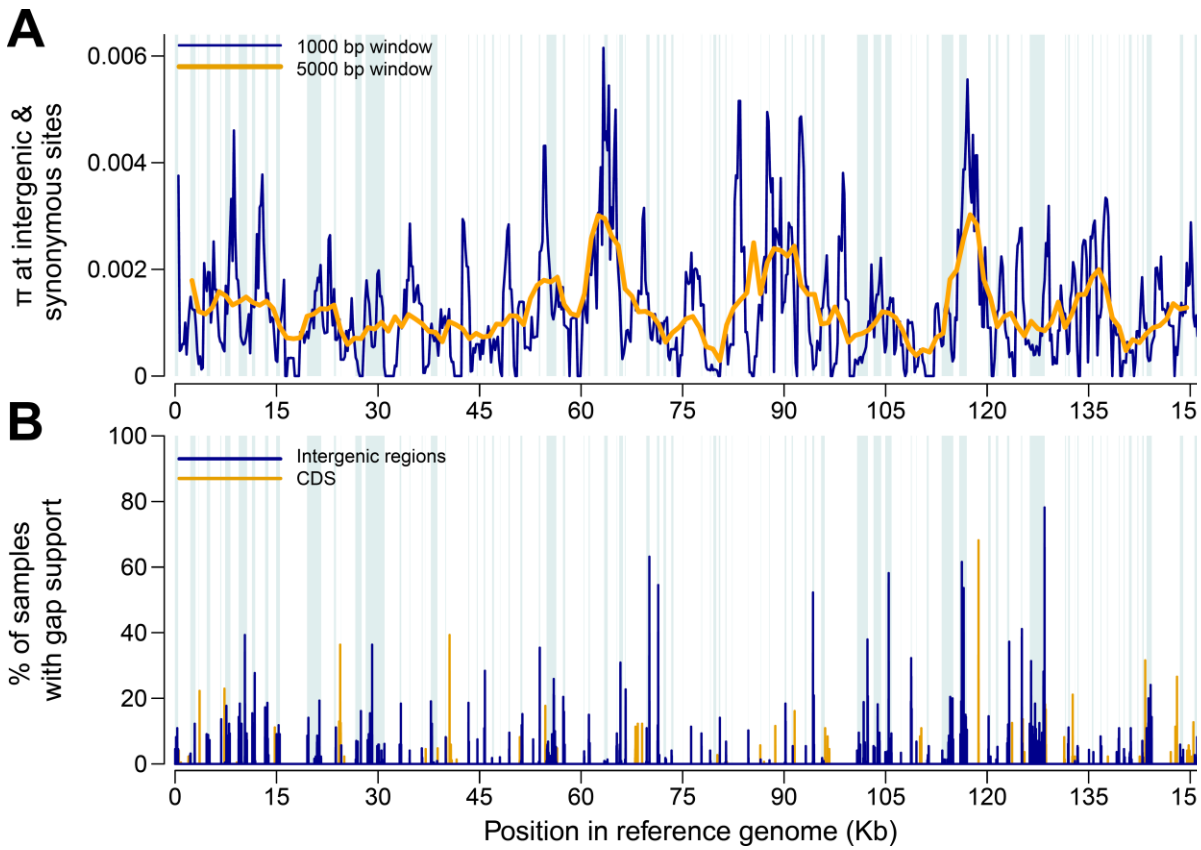


Fig. 3.4. **Neutral site-diversity (A) and support for short insertion deletion polymorphisms (B) across the Kallithea virus genome.** Figure shows, in panel A, variation in nucleotide diversity across non-coding and synonymous sites in the Kallithea virus genome, plotted as a sliding window with two window sizes. The lower panel (B) shows the percentage of Kallithea virus infected samples which showed evidence of a gap (representing a short insertion deletion polymorphism, with support of at least 5 reads) at each position in the reference genome. Intergenic regions of the genome are coloured in grey. I found a strong positive association between intergenic regions and finding InDels (X-squared = 3236, df = 1, p-value < 2.2e-16).

3.4 Discussion

I utilised a continent-wide pooled sequencing dataset to analyse spatiotemporal patterns of prevalence, and global and local levels of genetic diversity in the DNA viruses of European *Drosophila*. In comparison to *Drosophila*-infecting RNA viruses, the majority of DNA viruses appear to be extremely rare, meaning that only three of the viruses identified in Wallace *et al.* 2020 had sufficient samples for analysis. These three viruses – Kallithea virus, Vesanto virus and Linvill Road virus – showed differing patterns of prevalence, and genetic diversity. However, all viruses showed high population differentiation (F_{st}) and limited fine-scale predictability in virus incidence, suggesting that the dynamics of these *Drosophila* DNA viruses are characterised by transient epidemics and regular genetic bottlenecks.

3.4.1 Broad-scale spatiotemporal patterns of virus prevalence in DNA viruses of European *Drosophila*

All three viruses analysed in this study showed individual fly-level prevalence estimates of <3%, indicating that DNA viruses are rare in European *Drosophila melanogaster*, in comparison to some DNA viruses infecting Drosophila populations in the U.S.A (Unkless 2011). Kallithea and Linvill Road virus prevalence (see Fig. 3.1, panel A and B), was predictable in space, with similar, long ranges of spatial autocorrelation between samples (see Fig. 3.3). However, for all three viruses analysed, sampling site had no effect on the likelihood of infection. This suggests that, although DNA virus incidence may be more likely in a large geographic area, at a local level there is low predictability of virus prevalence. But should we actually predict local-level consistent prevalence of insect viruses? It's possible that so far, research has been biased towards higher prevalence viruses, and that for viruses with low prevalence, we should expect episodic, unpredictable bursts of infection when sampling isn't structured at a fine spatial scale.

I found an increased prevalence of Linvill Road virus, an ssDNA Densovirus, in samples with a higher relative percentage of reads mapping to *D. simulans*. This either points to these infections actually being in *D. simulans*, or that infections in *D. melanogaster* represent spillover. Previous metagenomic studies have found a high prevalence of Linvill Road virus in *D. simulans* (see data from Signor *et al.* 2018), supporting this hypothesis. This effect was robust to spatial autocorrelation (see Fig. 3.2), therefore a simple spatial correlation between more contaminations with Dsim in the South of the sampling area, where there is also more Linvill Road positive sites (Fig. 3.1 panel B), is not likely to be underlying this pattern. This alternative host species could also impact the evolution of this virus, and in particular, its effective population size. This would also explain Linvill Road virus fly-level prevalence being so low (0.34%, 0.21-0.5%) in a primarily Dmel dataset, and its fine-scale spatiotemporal unpredictability.

In contrast to Linvill road and Kallithea virus prevalence, which was not predictable in time, Vesanto virus, a recently extended Bidna-like virus, showed higher prevalence in collections from 2015 and 2016 in comparison to 2014 (Fig. 3.2). This lower prevalence in 2014 could be the reason why only a single segment of this Bidna-like virus was described in (Kapun et al. 2020), and its increase in subsequent years suggests that this virus could be increasing in prevalence over time. However, further sampling will be required to distinguish these year to year fluctuations from long–term trends. Luckily, the DrosEU consortium is continuing to sample *Drosophila* from this system, and further sequencing will provide just such an insight into the long-term dynamics of their DNA viruses, which could also be combined with sequencing across a wider spatial range (eg. Kapun *et al.* 2021). It also remains to be seen whether copy number variation within the segments of the Vesanto virus genome, a significant level of which was identified in Wallace *et al.* 2020, shows any spatiotemporal, or host species associated pattern. In this study I measured presence / absence of Vesanto virus as total Vesanto virus copy number across all segments. However, viral segment copy number can

and does differ in this dataset, possibly due to their role in expression regulation. Other multi-partite viruses also show evidence of variation in copy number between both populations, and host species (eg. Moreau *et al.* 2020), which could lead to different effective population sizes for each segment, and in turn influence their frequency of reassortment, rates of mutation and substitution, and rates of host switching (see review Michalakis & Blanc 2020). The spatiotemporal incidence of specific Vesanto virus segment haplotypes would also be interesting to quantify with further sampling, to see whether the higher incidence of Vesanto in later years of data is associated with increasing prevalence of a specific haplotype.

Collection season, and the presence of *Wolbachia* endosymbionts had an insignificant effect on prevalence for all three viruses. The absence of any detectable effect of season on prevalence at both a location specific and continent-wide scale is likely due to the low prevalence of these DNA viruses, though other microbiota of European Dmel also show a lack of seasonal effects (Wang *et al.* 2020). This low prevalence creates less power to detect seasonal effects, as quantitatively distinguishing between continuous low prevalence across space and time, and short episodic epidemics is much harder than identifying long-term spatiotemporal trends in high prevalence pathogens. Low prevalence also means that epidemic bursts of infection are less likely to happen in periods of high host abundance, such as the late collecting seasons in *Dmel* (Basden 1954). Older studies of Baculoviruses and Nudiviruses in other temperate forest insects predict epizootics when host densities are high, and that these viral infections play some role in regulating the population of their hosts, producing cyclic variations in abundance of the insect species if disease induced mortality is high and loss of infective particles from the environment is low (Anderson & May 1980). However, in the case of these three viruses I think it unlikely that they could drive such effects, given their low prevalence and fine-scale unpredictability.

3.4.2 An endogenous copy of Galbut virus likely had a Northern European origin and now shows seasonal patterns of incidence

In the genomes of arthropods, there are many examples of endogenous, integrated copies of RNA viruses (eg. Ballinger et al. 2012, and example of partiti-vius EVEs in Aguiar *et al.* 2020). I applied these methods of analysing the broad-scale spatiotemporal patterns of virus incidence to an endogenous copy of Galbut virus found segregating in the genome of European Dmel in Wallace et al. 2020. Galbut virus is an extremely common dsRNA partiti-virus of Dmel and Dsim, a copy of which integrated into the European Dmel genome in the last 300 years (Wallace et al. 2020). I found that the Galbut EVE had a patchier spatial distribution than the viruses analysed (see Fig. 3.3 for range of spatiotemporal autocorrelation), which was seasonally dependent (Fig. 3.1 panel C and D) and appears to be spatially limited to Northern Europe. This leads us to hypothesise that the original integration of this EVE happened in Northern European populations of Dmel. The shift of this EVE into the Nordic regions in the late sampling season is likely due to the expansion of Dmel range with rising temperature, but its decreased prevalence in Scotland, Poland and Ukraine is harder to explain. Seasonal changes have been found to significantly impact both the community composition of temperate populations of Dmel on both a phenotypic level (Behrman et al. 2015), and in underlying allele frequencies (eg. Cogni et al. 2014; Machado et al. 2019). Based on this, we might hypothesise that individuals carrying this EVE are more competitive at lower host abundance or lower temperatures, but it is also possible that this shift in distribution is simply a sampling artefact (only 7 pools in the late sampling season were positive for the Galbut EVE).

The amount of admixture between Northern and Southern European Dmel will influence the speed of the drift this EVE through populations. However, as population differentiation tends to be on a longitudinal, rather than latitudinal gradient in populations of European *Drosophila* (Kapun et al. 2020) it's unlikely that low admixture with southern populations has limited this EVE to Northern regions, and more likely that its limited spatial spread is simply due to its

recent insertion. If this EVE confers any kind of selective advantage on the Dmel host, such as increased immunity to other exogenous viruses, then this would also affect the speed of its spread through the population. In mosquitoes, non-retroviral EVEs have been found to often occur in pi-RNA generating clusters (Palatini *et al.* 2017b), and these piRNAs have been hypothesised to confer anti-viral protection on their hosts (Tassetto et al. 2019). However, in Wallace et al. 2020 the location of this EVE in the host genome was not identified, and so the fitness effect of this EVE remains unknown.

3.4.3 Genetic diversity is variable, but population differentiation high in Drosophila-infecting DNA viruses

DNA viruses are a middle ground between RNA viruses (small genomes, high mutation rates) and eukaryotes (larger genomes, low mutation rates). Because of this, we might expect the levels of genetic diversity, and patterns of adaptive evolution in the DNA viruses of *Drosophila*, to also be somewhat intermediate between their hosts and RNA viruses. However, in comparison to their hosts, the DNA viruses in this system are low in prevalence and are experiencing regular genetic bottlenecks at both a population and host scale, dampening their adaptive potential at a global population scale. I aimed to find out what patterns of diversity and evolution these conditions produce, and how this differs amongst different types of DNA viruses.

The mutation rates of DNA viruses vary depending on their single, or double stranded nature, and their genome size (Duffy 2018). Consistent with this, I found that diversity at synonymous sites varied between the different DNA viruses examined, due to their genome organisation, prevalence, and infection biology. I examined pairwise diversity (π), as this measure of divergence doesn't require a known population size. The double-stranded, large Nudivirus, Kallithea virus, showed very low diversity at synonymous sites (total π_S = 0.15%, local π_S = 0.04%) with little pattern to its variation across the genome (Fig. 3.4, A). This level of diversity

is almost ten-fold lower than that of its host (Kapun *et al.* 2020) likely due to Kallithea virus's regular genetic bottlenecks and low prevalence (estimated at 2.1%), which even the higher mutation rate of the virus (estimated ~10^{-7}-10^{-8} - Duffy 2018) cannot overcome. In contrast, single stranded DNA viruses show mutation rates only slightly lower than RNA viruses (Duffy 2018) resulting in relatively high levels of genetic diversity within and among other host species (eg. Ge *et al.* 2007). This data seems to reflect this, as diversity at synonymous sites in the single-stranded DNA viruses, Linvill Road and Vesanto virus, was comparably higher. In comparison to Kallithea virus, the synonymous site diversity of Vesanto virus was ~7 fold higher (total π_s = 1.16%, mean local π_s = 0.28%), and the synonymous site diversity of Linvill Road virus was ~9-5 fold higher (global π_S = 1.45%, local π_S = 0.21%). The slightly confusing aspect of this comparison is that Linvill Road virus shows almost 10 fold lower prevalence than the two other viruses, which we might expect to reduce its diversity. However, if, as suggested by the linear model results (discussed above), Linvill Road virus infections are maintained at a higher prevalence in *D. simulans*, this could explain its low prevalence's lack of effect on diversity.

Population differentiation is high for all three DNA viruses (F_{st} > 0.7), regardless of genome type, and far higher than that of the host population (differentiation roughly F_{ST} = 0.03 in the DrosEU samples collected in 2014, Kapun *et al.* 2020). For all three viruses, global pairwise diversity at neutral or nearly neutral sites is also consistently far greater than that at the local level and the vast majority of polymorphisms, both in single nucleotides and short indels, are globally rare and low in frequency. The limitation of the majority of these SNPs to 1 or 2 populations is likely due to many of the variants being deleterious and therefore selected against at the global population level, or having recently arisen. Many pools in this dataset also probably only contain a single infected fly, meaning that I was probably examining within host diversity in many of the samples, consistent with the low synonymous site diversity observed in this study. It is also possible that many of these numerous local SNPs represent sequencing errors, or were generated by the miss-assignment of reads due to barcode

switching (Ballenghien *et al.* 2017). However, this is unlikely given the conservative base quality threshold applied (Li *et al.* 2009), and the use of a 1% minimum allele frequency.

3.4.4 Patterns of evolution in DNA virus genomes

For Kallithea virus, I examined which SNPs were maintained across samples, and found, as expected, that the majority of the 28 SNPs which were present in at least a third of samples were in intergenic regions or were synonymous, also consistent with selection against deleterious, non-synonymous SNPs. Only two non-synonymous SNPs occurred in at least a third of samples. One was in the DNA Helicase 2-like protein (AQN78620.1), of interest because the Helicase gene, which unwinds DNA and is critical in its repair, is one of the most rapidly evolving genes across Baculoviruses and Nudiviruses (Hill & Unckless 2017, 2018a). Additionally, a SNP in the Helicase-2 gene was recently found to contribute to the evolution of a high titre variant of the close relative of Kallithea, Drosophila innubila Nudivirus (Hill & Unckless 2020).

The different viral genome organisation, and prevalence, of the three viruses also appears to cause variation in their evolutionary constraint – inferred from their ratios of non-synonymous to synonymous diversity on whole genomes, and individual genes. For Kallithea virus, constraint on evolution was low and several genes of interest showed a $\pi_A/\pi_S > 1$ indicating diversion from neutrality. SNPs in homologs of two of these genes (odv-e56-like protein (AQN78623.1) and DNA helicase-2-like protein (AQN78620.1)) have been recently found in the high titre variant of DiNV (Hill & Unckless 2020), and these genes were also found in windows in the lowest 2.5 percentile of Tajima's D, indicating recent selection (Hill & Unckless 2018b). Another of the Kallithea virus genes with a $\pi_A/\pi_S > 1$ was a trypsin-like serine protease-like protein (AQN78562.1). Trypsin-serine like proteases can be both upstream and downstream of key immune proteins like toll, so possibly evolution of a Trypsin-serine like protease in the virus could interfere with the immune response of the fly. A trypsin-serine like

protease containing region of the DiNV genome also showed evidence of recent evolution (Hill & Unckless 2018b).

However, a binomial test revealed no genes showing significant evidence of positive selection in their ratio of non-synonymous to synonymous SNPs. This is consistent with Kallithea viruses' close relatives, which show little evidence of adaptive evolution in their genomes (Hill & Unckless 2018b). Consistent with this genome wide adaptive constraint, I found 9 genes which showed significant evidence of negative/purifying selection, including two lef genes, essential for replication, and gp83, a suppressor of Toll activity (Palmer *et al.* 2018a). The single-stranded DNA viruses displayed higher levels of whole genome evolutionary constraint in comparison to Kallithea virus. This could be due to their smaller segment/genome size, which limits the mutational load they can sustain (Duffy *et al.* 2008).

3.5 Conclusions

I used a continent-wide, multi-year dataset of pooled sequencing samples to characterise the broad scale variation in prevalence, and genetic diversity, of DNA viruses in *Drosophila melanogaster*. Although I find broad-scale trends in virus prevalence, spatially for Kallithea and Linvill Road virus, and temporally for Vesanto virus, the spatiotemporal predictability of these viruses at a fine-scale remains low. A more spatially dense distribution of samples, and a longer timescale of sampling could help to differentiate between spatiotemporally constant low prevalence, and bursts of higher virus prevalence with predictability at the local scale, and more effectively characterise this fine-scale variation in prevalence. I found diversity at synonymous sites varied between the different DNA viruses examined, due to their genome organisation, prevalence, and infection biology. However, I consistently found high levels of population differentiation (F_{st}), suggesting that DNA virus populations undergo regular genetic bottlenecks and that their dynamics are characterised by irregular, and local-scale epidemics. In the future, more insect-virus studies in wild systems would benefit from the ability to quantify the level of fine and broad scale spatiotemporal predictability in virus incidence (though see Myer & Johnston 2019), as we have tried to do on a broad-scale sampling scheme in this study.

4 Viruses transmitted from wild *Drosophilidae* can reduce offspring number and lifespan in *Drosophila melanogaster*

Wild Drosophila collected in the South of England for this experiment were collected by Darren Obbard. All other wild collections, experimental fly and lab work was carried out by MW, after initial instruction on wet lab protocols was given by Nathan Medd. Initial MCMCglmm models were written with the help of Darren Obbard and Jarrod Hadfield, and were refined by MW.

Data availability – Data, and code for models and figures in this chapter can be found in a GitHub repository at https://github.com/megan-a-wallace/fitness_effects

4.1 Introduction

In the wild, insects and their viruses are assumed to be involved in reciprocal, co-evolutionary arms-race dynamics. This assumption is supported by the fast evolution of insect immune genes (Obbard *et al.* 2006), the presence of large-effect polymorphisms for resistance in insect genomes (eg. Cao *et al.* 2016), and the evolution of viral suppressors of the host immune response (eg. Palmer *et al.* 2019). However, the strength of virus-driven selection on host immune genes is dependent on the cost of viral infection to host fitness, and we have little idea of the range of fitness effects imposed on insect hosts by their full complement of naturally-occurring viruses. We can use the model organism *Drosophila melanogaster* (Dmel), other species of *Drosophila,* and their naturally-occurring viruses to assess the possible strength of virus-driven selection in wild insects, by assessing the fitness costs of wild *Drosophila* viruses.

In the wild, and in the lab, Dmel individuals are infected with a range of viruses described via pathological, classical virology, and metagenomic sequencing methods. In the wider genus

Drosophila, over 130 diverse *Drosophila*-associated viruses have now been described (Unkless 2011; Webster *et al.* 2015, 2016; Medd *et al.* 2018). Eight of these *Drosophila* viruses have been isolated in the lab, usually the first step to characterising fitness effects. However, as they encompass <10% of the *Drosophila* viruses described, relying solely on isolated viruses in experiments is likely to provide us with an incomplete picture of the strength of virus-driven selection.

These isolates come from six RNA viruses: Dmel Sigma virus (DmelSV), *Drosophila* C virus (DCV), *Drosophila* A virus (DAV), Dmel Nora virus (DmelNV), La Jolla virus and Galbut virus (Berkaloff *et al.* 1965; Jousset *et al.* 1972; Plus *et al.* 1975; Habayeb *et al.* 2006; Carrau *et al.* 2018; Cross *et al.* 2020), and two DNA viruses: *Drosophila innubila* Nudivirus (DiNV) and Kallithea virus (KV) (Unkless 2011; Palmer *et al.* 2018b). The fitness effects of these infections range from severe to more subtle pathologies. For example, injection with one of the best characterised *Drosophila* viruses, DCV, causes impaired locomotor activity, metabolic depression, and eventual mortality in Dmel adults (Arnold *et al.* 2013a). These phenotypes result from DCVs infection of foregut tissues, specifically muscular tissue surrounding the crop, and the resulting intestinal obstruction (Chtarbanova *et al.* 2014). Additionally, both isolated DNA viruses of *Drosophila* (DiNV and KV) incur severe fitness costs on their hosts (Unkless 2011; Palmer *et al.* 2018b). Systemic injection with DiNV (dsDNA, *Nudiviridae*) causes a decrease in lifespan in both *D. innubila* and *D. falleni*, and infection in wild females of both species incurs an ~80% reduction in offspring production (Unkless 2011). Interestingly, this study directly observed a virus-induced reduction in female fecundity, finding a significant reduction in ovariole number in infected flies. Systemic injection of KV, another large dsDNA Nudivirus, into Dmel causes increased male mortality, and for females, decreased movement and late-life egg laying (Palmer *et al.* 2018).

Other viruses have more subtle pathologies, such as the vertically transmitted DmelSV, which causes no detectable decrease in adult mortality but does reduce fitness by an estimated ~25% in some experiments (Yampolsky *et al.* 1999; Wilfert & Jiggins 2013). Persistent

infection with DmelNV (+ssRNA, *unclassified Picorna-like*) also produces no obvious pathologies. Habayeb et al. (2009) found that the survival of Dmel infected flies was similar to that of uninfected flies up to 50 days. However, both wild-type and white-eyed (w^{1118}) Dmel infected with DmelNV faecal-orally show differential expression of genes from several immune-related pathways (Cordes *et al.* 2013; Lopez *et al.* 2018). For DAV (+ssRNA), commonly described as the most benign of the isolated *Drosophila* viruses, again no apparent pathologies have been found on injection into Dmel, despite isolation allowing extensive characterisation of its genome and virion structure, and its rampant replication in lab flies and cell culture (Christian 1987; Ambrose *et al.* 2009). However, DAV reduces lifespan on intra-thoracic injection into the soft fruit pest species *Drosophila suzukii* (Carrau *et al.* 2018), suggesting that the virulence of *Drosophila* viruses commonly varies across host species (eg. Longdon *et al.* 2015; Beraldo 2018 - unpublished thesis data).

Species specific metagenomic sequencing of *Drosophila* has also revealed overlaps in host range for many *Drosophila* infecting RNA viruses (Webster *et al.* 2015, 2016; Medd *et al.* 2018a). This suggests that in the wild, the cross-species transmission of some *Drosophila* infecting RNA viruses may be common (supported by data from Ch2 of this thesis) and creates an opportunity to study arthropod viruses in a multi-host multi-virus community. However, the route by which most of these viruses are transmitted between individuals, or species, remains unknown.

Characterising the real fitness costs of *Drosophila*-virus infection in the wild is also hampered by the reliance of experiments on injection into the thorax or abdomen, transmission routes which are thought to be rare in the wild (see methods in Merkling & van Rij 2015). These methods allow more control over infection, and sample size, in the lab, but are unlikely to replicate wild fitness costs. Studies which use oral delivery of viruses, an infection route more common in the wild, have found several key differences in their infection phenotypes in comparison to studies using injection. For example, faecal-oral virus delivery can result in the infection of different tissues when compared with injection into the body cavity (Ekström &

Hultmark 2016). This tissue tropism based on infection route can also create different infection outcomes and immune responses (see review Mondotte & Saleh 2018). For example, DCV injected into adult flies induces mortality over 4-8 days, and even produces reductions in locomotor activity at sub-lethal viral titres (Hedges & Johnson 2008; Gupta *et al.* 2017a). In contrast, when DCV is fed to adult flies the same sub-lethal dose has no significant effect on locomotor activity (Gupta *et al.* 2017a). Oral infection of DCV and Dmel Nora virus also leads to the induction of elements of the Toll pathway, a phenotype which is not induced on injection of either virus (Ferreira *et al.* 2014). Only a couple of studies on DiNV and DmelSV (Yampolsky *et al.* 1999; Unkless 2011) have characterised fitness effects in the wild directly. Using this method makes it difficult to ensure sufficient statistical power to detect effects, but, for large-effect viruses, gives a much more accurate picture of the strength of virus-driven selection in natural populations. Studies which use individuals, and viruses from the wild, with the natural infection routes of insect pathogens, will give a more realistic estimate of evolutionary trajectory of viruses in wild populations.

The small number of isolated and characterised *Drosophila* viruses mean that the pathology, transmission route, and fitness costs of most *Drosophila* viruses remains unknown. In this study I used a more natural transmission route - contact between infected and uninfected individuals - to infect Dmel females with a subset of their native viruses, and viruses carried by other species of *Drosophila*. In the infected females, I assessed the effect of viral infection on two key components of fitness: lifespan and offspring production. Any reductions to these key components of individual fitness could create significant selection pressure on the viral hosts, and potentially drive the evolution of host immune genes.

4.2 Methods

4.2.1 Exposure of Dmel to wild Drosophila

To detect fitness costs associated with natural virus infection, we exposed laboratory flies to viruses carried by wild-collected flies, and followed their subsequent lifespan and offspring production. A summary of the methods used to expose Dmel OreR to viruses and record fitness-associated traits can be found in Fig. 4.1**.** First, to transmit or 'donate' naturally infecting *Drosophila* viruses which might be able to infect laboratory lines of Dmel, wild *Drosophila* were collected from three species groups. From the 2-8th July 2017 aspiration and netting from banana and yeast baited traps in Sussex (51.100N, 0.164E) and Edinburgh (55.919N, -3.211W) was used to collect wild Dmel, *Drosophila immigrans* (Dimm), and species from the *Drosophila obscura* species group: *Drosophila obscura* (Dobs), *subobscura* (Dsub), and *subsilvestris* (Dsus). All references to Dobs refer to the combined *obscura* species, which I combined into one treatment group because the three species are difficult to distinguish morphologically, and carry a similar range of viruses in the wild (Webster *et al.* 2016). To create high density agar vials, potentially containing lots of shed viruses from the wild flies, I separated the wild collected flies to species by morphology, and housed them on agar medium for 2-7 days.

To provide stocks of experimental Dmel to be exposed to wild viruses, I set up bottles of age matched OregonR *Wolbachia* negative flies (OreR Wol-) (inbred line) using eggs laid onto agar and 'squirted' on to Lewis medium. This method of stock creation, where eggs are laid onto agar medium in a large cage, before being suspended in PBS and a specific volume squirted into new Lewis vials, limits the faecal-oral spread of viral infection between adult *Drosophila* and larvae, and controls the density of flies in each vial. One day after eclosion, OreR from these bottles were separated into agar vials (25 females and 10 males) to mate, but not lay eggs. Samples of 10 OreR males were also frozen from each of the recipient stock bottles to test them for any contaminating viral infections. After four days, 20 OreR females

were added to each one of the vials containing the species of wild *Drosophila* (Dobs, Dimm and Dmel), so that the exposure vials contained 10 wild flies of one species and 20 OreR Dmel females. To control for other effects of being exposed to wild species, independent of viral infections, 20 OreR females were also co-housed with 10 OreR males in 5 agar vials, at the same vial density as those co-housed with wild flies.

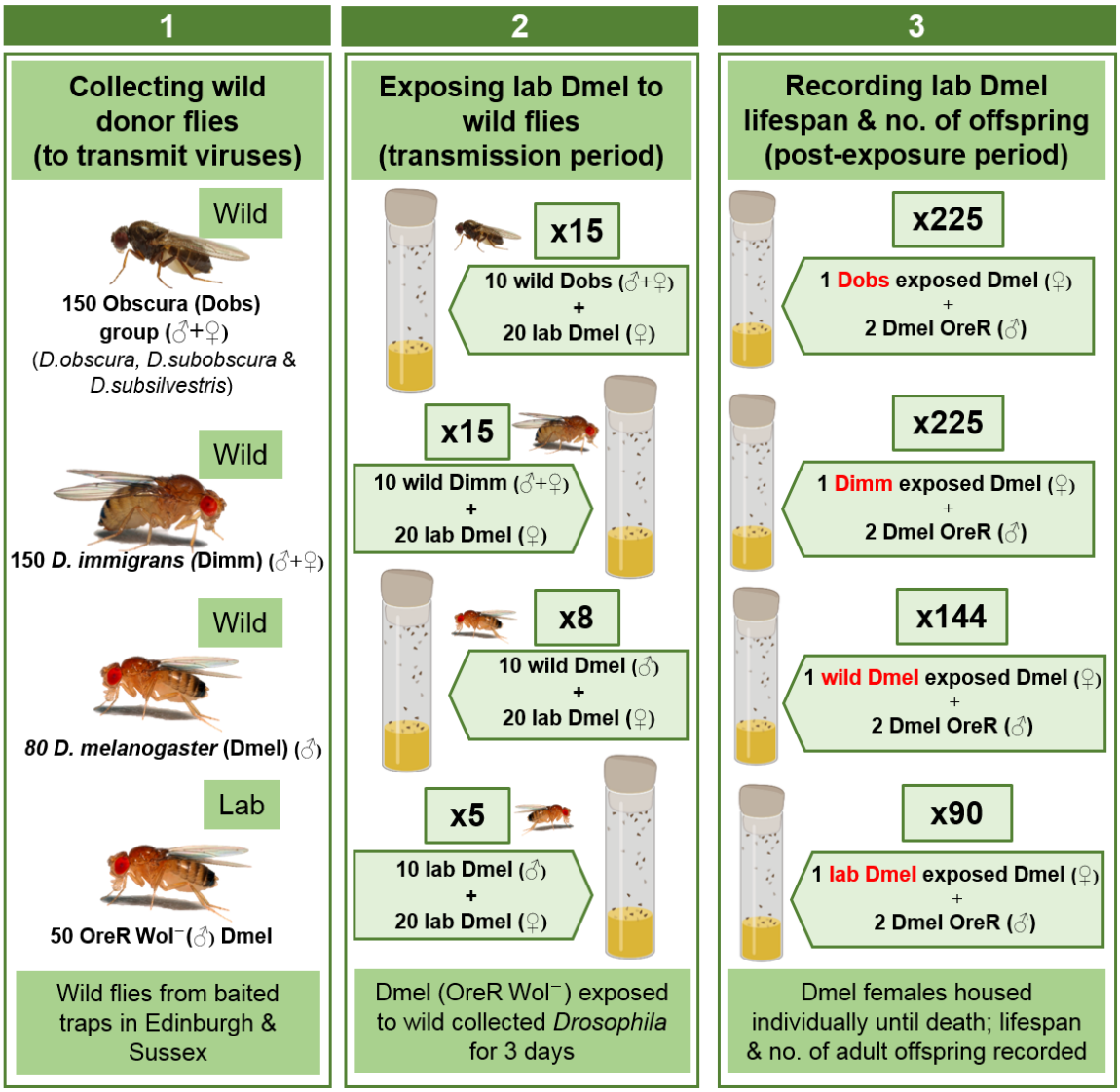


Fig. 4.1. Summary of methods used to expose female Dmel to naturally-occurring Drosophila viruses. Summary of the experimental design used to; (1) collect wild Drosophila of three species groups to transmit viruses, (2) expose Oregon R *Wolbachia* negative Dmel females to these wild collected species, and (3) record the key components of the exposed Dmel female's fitness. *Drosophila* photos taken by Darren Obbard.

All the co-housing vials (wild species and control vials) were left on a lab bench for three days to allow viral transmission between the wild flies and lab females. To find out what viruses the female OreR had been exposed to during the three days of co-housing, the wild donors (or OreR lab males in the control vials) were removed after this period and frozen for later virus RT-PCR assays. To allow us to distinguish between laboratory and wild-collected *D. melanogaster*, all lab 'recipient' flies were female, and all wild 'donor' flies were male. From each co-housing vial, the female OreR were then placed into individual vials containing Lewis medium for the remainder of their lives. To control for possible differences in mating stress, 2 male OreR were added to all of the vials containing individually housed females.

On a daily basis, any dead OreR females were removed and frozen for later virus assays, their lifespan recorded, and any dead males replaced to keep stress from courtship or mating at a constant level across vials. As a proxy for fecundity, I recorded the number of emerged adult offspring produced by each co-housed female 15 days after tipping. Emerged offspring numbers were recorded more frequently during the first two weeks of the experiment – at days 2, 5 and 10 – after which the female OreR were tipped every 7 days. The experimental design was such that the only difference between normal laboratory maintenance of OreR and the treatment of the female OreR in this experiment was the 3 days they spent exposed to wild *Drosophila,* during which time we expected them to be exposed to, and catch, viruses.

4.2.2 Detection of Drosophila virus infection & transmission events

RNA was extracted from the wild donor flies and single OreR Wol- flies using a manual Phenol-Chloroform protocol. The flies were macerated in TRIzol reagent (ThermoFisher Scientific) (100µL for single flies, 200uL for pools of 10) and RNA extracted using the manufacturer's instructions. cDNA from each sample was synthesised using random hexamer primers (ThermoFisher Scientific) (4µL 10 mM random primers per 1µL total RNA) which were then incubated with the RNA (70ºC, 5 minutes). RNase free H_2O, 10 mM mixed dNTPs

(ThermoFisher Scientific), M-MVL reverse transcriptase (200 units/µl, Promega) and 5x M-MVL reaction buffer (Promega) were then added to the 5µL RNA-random primer mix and incubated at 37ºC for 60 minutes.

To reveal the viruses to which the female OreR were exposed during co-housing, I tested each of the wild collected donor groups of *Drosophila* (Dimm, Dobs and Dmel) and the control OreR males for a total of 59 different viruses by RT-PCR (primers listed in table S4.1). Five of the viruses (Vogrie virus, Crammond virus, Sighthill virus, Sunshine virus and Burdiehouse burn virus) are among those newly described in chapter 2. To find out which viruses had transmitted to the Dmel OreR, the recipient females were assayed for those viruses to which they were exposed during co-housing. Because the OreR females were taken from their individual vials after death, much of the RNA extracted from these recipients was severely degraded. This required the design and use of additional PCR assays with relatively short product lengths (120-170bp). Details of the short product primer assays can be found in Table S4.2.

All PCRs were performed with a master mix of the following volumes per 1µl of cDNA template - 1µl 10xNH4 buffer, 0.25µl 50mM MgCl2, 0.05µl 5U/µl BIOtaq DNA polymerase (all Bioline reagents ltd.), 0.3µl 10mM mixed dNTPs (Life Technologies), 7.5µl RNase free H2O, and 0.5µl of both 10µM forward and reverse primers (Sigma Aldrich). All PCRs were run on a thermocycling regime of 94ºC for 5 minutes, then 5-10 x (94ºC for 15s, * ºC for 30s, 72 ºC for 1 minute) dropping 1^0C every cycle, then 25 x (94ºC for 15s, * ºC for 30s, 72 ºC for 1 minute), then 72 ºC for 5 minutes, with the * ºC adjusted for virus assays. The Tm for each primer pair can be found in table S4.1/table S4.2. Presence or absence of a virus was determined using gel electrophoresis, stained using GelRed 10,000x (Biotium), on a 1 or 2% agarose gel (dependant on the size of the PCR products) at 90V.

For each wild collected 'donor' species, I recorded the viruses present in each of the co-housing vials, and which viruses transmitted to the OreR Dmel females. As the wild individuals were combined by exposure vial for virus assays, the number of wild flies infected in each pool

is unknown. Therefore, it is not possible to calculate precise individual-based wild prevalence estimates from this data. However, I calculated approximate estimates of global virus prevalence in the wild host species, using the proportion of donor species vials infected with the virus. To calculate these figures, I used a maximum-likelihood function based on a model of pooled Bernoulli trials (eg. Speybroeck *et al.* 2012 - as described in Ch2 and 3). I then quantified transmission for each specific host: virus combination by calculating the number of individuals infected with a virus as a percentage of the number exposed. One of the control vials of OreR males tested positive for Muthill virus, and two for DimmNV, and any transmissions resulting from these exposures were analysed in the same way as the wild exposures.

4.2.3 Sanger sequencing of transmitted viruses

DimmNV has not previously been reported infecting Dmel (van Mierlo *et al.* 2014), and DimmSV is expected to have purely vertical transmission (Longdon *et al.* 2017), making any observations of either virus in Dmel surprising. Therefore I used Sanger sequencing to see whether the viral infections seen in the donors and recipients in this experiment lie within the currently understood phylogeny of DimmNV, and DimmSV, and whether I was observing a host shift into Dmel, or whether I was cross-priming on to another related virus. I Sanger sequenced a ~600bp region of the putative RNA-dependent RNA polymerase (RdRp), Viral protein 2 (VP2), of DimmNV. I also Sanger sequenced regions of the L (~630bp) and N (~525bp) genes of DimmSV. I sequenced all viral PCR products from both the wild collected donor flies, and the recipient OreR Dmel females, and used the PCR reagents and volumes described above.

The products of successful PCR reactions were cleaned using Exo-SAP-IT ® Express Reagent (Applied Biosystems) according to the manufacturer's instructions. 4µL (recipients) or 1µL (donors) of the cleaned product were then sequenced by adding 1µL BigDye reagent from BigDye ® Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems), 0.7 µL 3.2mM

forward primers and RNase free H_2O (volume to make the total reaction 10 µL) and incubating for 1 minute at 96ºC, then 25x (10s at 96ºC, 5s at 50 ºC, 4 minutes at 60 ºC). Sequencing was then conducted at Edinburgh Genomics. The resulting sequences were quality trimmed at ends, and pools with multiple flies were checked for heterozygous sites using FinchTV (Geospiza 2004). A BLAST search then identified the closest virus sequence in NCBI databases (Altschul 1990). The viral sequences, additional sequences from other studies, and sequences from wild collected flies generated in chapter 2 were then aligned using a codon model in PRANK multiple sequence aligner (Löytynoja 2014). These alignments were used to construct phylogenies of the virus sequences, whilst inferring the most likely host species on ancestral nodes, in BEAST v1.10.4 (Rambaut *et al.* 2018). I used the SDR06 substitution model (Shapiro *et al.* 2006), and default priors, other than a lognormal prior with a mean and initial value of $5x10^{-4}$, and standard deviation of $2x10^{-4}$ (a reasonable rate of RNA viruses) on the strict clock rate. Estimates of population growth rate in preliminary models showed that a constant population size tree coalescent was most appropriate. For the DimmSV trees, I allowed ancestrally reconstructed host species transmission rates to vary independently, so that the rate of transmission from Dimm to Dmel was allowed to differ from the reverse direction of transmission. Two Markov chain Monte Carlo (MCMC) chains with a length $5x10^7$ were run, sampling 10,000 trees on each chain. I examined trace files for sufficient effective sample size and chain convergence before combining runs and generating a maximum clade credibility (MCC) tree in TreeAnnotator v1.10.4 (Drummond & Rambaut 2007), setting the burn-in to 10% of the total MCMC chain length. The MCC trees were viewed in FigTree v1.4.4 (http://tree.bio.ed.ac.uk/) and then annotated in R v.4.0.3 (R Core Team 2020) using the ggtree package (Yu 2020).

4.2.4 The impact of viral infection on Dmel lifespan

To find out whether viral infections can reduce, or even increase, lifespan I used a linear mixed effects model. I estimated the posterior distribution of the parameters using a Bayesian approach, implemented in the R package MCMCglmm (Hadfield 2010). The response variable

of lifespan was modelled using a Gaussian distribution. The final model included each transmitted virus as a fixed effect (10 factors with two levels: infected and uninfected) and exposure vial as a random effect to control for exposure species. The sample size of the Markov chain was 10,000, and all parameter estimates had an effective sample size of > 6,000. The formula of this model was;

Dmel lifespan ~ Chaq satellite + Galbut virus + Muthill virus+ DimmNV +DmelNV+ DimmSV +Grom virus + Prestney burn virus + Corseley virus + Larkfield virus + ~f(experimental vial)

I compared this model to a second linear mixed effects model in which the type of viral infection was specified as a random effect, which varied around the effect of being infected by any virus. The number of different viruses infecting a host was included as a fixed effect (continuous). This model allowed me to ask whether the virus species caused a significant deviation in lifespan from the general 'virus effect', and whether viral infections have an additive effect on lifespan when co-infecting individuals. Exposure vial was again included as random effect in the model to account for variation caused by exposure species and vial. The formula for this model was;

Dmel lifespan ~ no. of viral infections + idh(at.level(infected)):(all virus species) + ~f(experimental vial)

The sample size of the chain was 10,000, and all estimates from this model had an effective sample size of >9,500. I chose to focus on the 'virus as fixed effects' model due to its greater simplicity of interpretation.

I had insufficient power to detect fitness costs for many of the transmitted viruses in this experiment. To assess what sample size would be needed in future experiments to detect different strengths of virus effects on lifespan, I ran power simulations using the R package simglm (LeBeau 2020). In these simulations I varied sample size to a maximum of 100,000 flies per treatment, and the mean reduction in lifespan caused by a virus between 1% and

20%. These simulations assume a mean (41.9) and variance (266.7) in lifespan as measured from this experiment, and a balanced experimental design with only one viral infection.

It is possible that flies may clear infections prior to death, causing some infections to be missed (eg. in DCV, Mondotte et al. 2018). To test the impact of this on my inferences, I selected one common virus (DmelNV) and additionally tested males and offspring, who may have received the virus from an infected female who later cleared the infection. The males housed with the virus exposed females were frozen on death of the female in their vial, and so before their own death. I also froze offspring produced in the first 2 days of the experiment. Therefore, by taking not only female infections, but also male and offspring infections into account, I picked up DmelNV infections I might have missed in these cleared flies, or because of degraded RNA. I then included the male and offspring infections in the data by assigning a female Dmel individual as infected with DmelNV if either she, her accompanier males, or her offspring were infected (16 additional infections). I analysed this data using the 'viruses as fixed effects' model, and all estimates had an effective sample size >5,500.

4.2.5 The impact of viral infection on Dmel offspring number

To find out whether viral infection can alter the number of offspring produced by female Dmel, I again used a linear mixed effects model. Lifetime offspring production was modelled using a hurdle-Poisson distribution. This model creates two latent variables associated with each data point; 1) the mean of a zero-truncated Poisson distribution and 2) the probability of zero offspring production (on the logit scale). Like the lifespan models, this model again included ten viral fixed effects (each a factor with two levels) and a random effect of exposure vial. However, the probability of zero-offspring was only allowed to vary in relation to eight of the viral infections, as no zero-offspring individuals were infected with DimmSV or Corseley virus, and therefore this probability was inestimable. The overall intercept of the model was

suppressed so that the estimates for the probability of zero offspring production could be interpreted independently of the Poisson model estimates. The formula for this model was;

Dmel offspring production ~ trait -1 + at.level(non-zeros):(all virus species) +
at.level(zeros):(estimable virus species)+ ~idh(non-zeros):(experimental vial)

Simulated samples from the estimated model fit showed that this model accurately predicted the number of zero values in the data (Fig. S4.2).The sample size of the chain was 200,000, and all estimates had an effective sample size of > 1,500.

A reduction in total offspring number might simply be a consequence of reduced lifespan, but could also be due to a lower rate of egg-laying, even in early life. To distinguish between these two possibilities, I also examined the effect of virus infection on a subset of the offspring production data, which only included the number of offspring produced until day 5 of the experiment (when all females were 13 days old). I excluded 10 flies which died in this period, and so lifespan has no effect on this data. Offspring production before day 5 of the experiment was significantly correlated with lifetime offspring production (Pearson's product moment correlation = 0.551, 95% CI = 0.493–0.604, t = 16.109, p = < 2.2e-16). I again used a hurdle-Poisson distribution for the response variable of 'offspring produced to day 5', which accurately predicted the number of zero-offspring females (29 in this subset of the data). I used the same fixed effects (the 10 transmitted viruses) and random effect (experimental vial) in this model as in the model of lifetime offspring production (see formula above). However, the probability of zero offspring was only allowed to vary in relation to five of the viral infections, as the probability of zero for the other five viruses (Grom, Prestney burn, Larkfield, DimmSV and Corseley virus) was inestimable, as no zero-offspring individuals were infected with these viruses. The sample size of the chain was 200,000 and all estimates had an effective sample size of > 4,000.

	Virus	Fraction of exposure vials carrying the virus	Estimate of wild prevalence (2-log likelihood bounds)	Estimate of transmission prevalence into Dmel (No. infected/no. exposed)*100
***Obscura* group species**	Buckhurst virus	5/15	4.0% (2.6 - 12.4)	DNT
	Withyham virus	3/15	2.2% (1.7 - 7.9)	DNT
	Pow Burn virus	1/15	0.7 % (0.7 - 3.8)	DNT
	Grom virus	7/15	6.1% (3.5 - 17.8)	4.71%
	Prestney burn virus	15/15	100% (78.9 - 100)	9.47%
	Lye green virus	9/15	8.8 % (4.6 - 24.6)	DNT
	Larkfield virus	15/15	100% (78.89 - 100)	2.11%
	Eccles virus	4/15	3.1% (2.1 - 10.1)	DNT
	Blackford virus	1/15	0.7% (0.7 - 3.8)	DNT
	Corseley virus	1/15	0.7% (0.7 - 3.8)	15.38%
	Sighthill virus	5/15	4.0% (2.6 - 12.5)	DNT
	Vogrie virus	4/15	3. 1% (2.1 - 10.1)	DNT
	Burdiehouse Burn virus	2/15	1.4% (1.2 - 5.8)	DNT
	Crammond virus	1/15	0.7% (0.7 - 3.8)	DNT
	D. immigrans Nora virus	1/15	0.7% (0.7 - 3.8)	DNT
	Muthill virus	1/15	0.7 % (0.7 - 3.8)	DNT
D. immigrans	*D. immigrans* Sigma virus	13/15	18.3% (8.6 - 49.8)	3.43%
	Bunya 4 / Sunshine virus	3/15	2.2% (1.7 - 7.9)	DNT

	D. immigrans Nora virus	14/15	23.72% (11.36 - 66.74)	7.91%
	Muthill virus	13/15	18.25% (8.61 - 49.85)	68%
D. melanogaster	Muthill virus	1/8	1.33% (1.26 - 7.18)	38.89%
	D. immigrans Nora virus	4/8	9.34% (6.05 - 29.44)	18.37%
	D. melanogaster sigma virus	2/8	2.84% (2.38 - 11.55)	**DNT**
	Torrey pines virus	1/8	1.33% (1.26 - 7.18)	**DNT**
	Craigie's hill virus	1/8	1.33% (1.26 - 7.18)	**DNT**
	Chaq*	8/8	97.67% (83.67 - 100)	7.35%
	Galbut virus	8/8	97.67% (83.67 - 100)	8.82%
	D.melanogaster Nora virus	7/8	18.77% (11.16 - 57.91)	37.50%
***D .melanogaster* (OreR controls)**	Muthill virus	1/5	2.21% (0.12 - 9.63)	29.41%
	D. immigrans Nora virus	2/5	4.98% (0.8 - 15.17)	6.06%

Table 4.1 Summary of viral exposure and transmission from wild flies to OreR Dmel. For viruses detected in the wild collected flies by RT-PCR, an estimate of the wild prevalence (and 2 log Likelihood bounds on this estimate) of these viruses was calculated in each wild species. 'Transmission prevalence' was calculated from the number of individuals infected with a virus as a percentage of the number exposed. DNT = Did not transmit (to Dmel females). *Chaq is thought to be either an optional segment, or satellite virus of Galbut virus (Cross et al. 2020).

4.3 Results

4.3.1 Transmission of Drosophila infecting viruses to D. melanogaster

4.3.1.1 Host range & prevalence of wild Drosophila viruses

PCR assays show that the wild collected Dimm, Dobs and Dmel co-housed with female lab Dmel were infected with 23 different viruses. Table 4.1 shows the viruses carried by each of the wild species, and which viruses transmitted to the OreR females. Only two viruses were found in all three species groups: D. *immigrans* Nora virus (DimmNV), a +ssRNA *picorna*-like virus and Muthill virus, a +ssRNA *Nege*-like virus. All other viruses occurred in only one wild species group. Estimated wild viral prevalence ranges between a maximum of 100% (2*lnL = 83.67-100%) for Galbut virus in Dmel, consistent with its high prevalence in the wild in Webster *et al.* (2015), and a minimum of 0.69% (2*lnL = 0.7-3.8%), for Blackford, Corseley, Crammond, DimmNV, Muthill and Pow Burn virus in Dobs (Table 4.1).

4.3.1.2 Transmission of wild Drosophila viruses to Dmel females

Nine viruses (Galbut, DmelNV, DimmNV, DimmSV, Grom, Prestney burn, Larkfield and Corseley virus) transmitted to the Dmel OreR females co-housed with the wild flies. The Dmel OreR females exposed to all three species groups contracted viruses, though a greater proportion of the viruses carried by Dimm and Dmel were transmitted to the Dmel OreR females (3/4 viruses from Dimm, 4/7 from Dmel, 4/16 from Dobs). Seven viruses – Grom, Prestney burn, Larkfield, Corseley, DimmSV, DimmNV and Muthill virus – were able to transmit across species into Dmel. Some host species: virus combinations produced a high percentage of transmissions. For example, for the Dmel exposed to Dimm carrying Muthill virus, 68% tested positive for the virus at the time of their death, suggesting that they acquired the infection and that it persisted for more than two days (mean lifespan of Muthill infected flies 38.4 days). Chaq, which is thought to be either an optional segment or satellite of Galbut virus, (Cross *et al.* 2020) was transmitted with Galbut virus in 10/12 cases and will be referred

to as Chaq satellite. The vast majority of viruses transmitted to <15% of the exposed individuals – meaning that for many viruses, I had relatively low power to detect the effect of viral infection on fitness.

DimmNV has not previously been observed to infect Dmel (van Mierlo *et al.* 2014). However, in agreement with data from chapter 2, I observed DimmNV infections in the wild collected Dmel, and transmission of this virus from wild Dmel and Dimm into 26 Dmel OreR individuals (transmission prevalence: Dmel – 18.37%, Dimm – 7.91%). I sought to confirm this transmission using Sanger sequencing of a ~600bp region of the DimmNV RdRp and comparison to known DimmNV sequences in NCBI databases. The DimmNV PCR products from two groups of wild collected donor flies (1 Dimm and 1 Dmel), and a recipient (infected from exposure to Dmel) were >99.5% identical to VP2 (RdRp) of DimmNV (KF242511.1 - the consensus sequence of an infected pool of 498 wild-caught flies from the UK as a part of van Mierlo *et al.* 2014). I constructed a phylogeny of this region of the DimmNV RdRp, including 10 DimmNV sequences from earlier chapters of this thesis, and a DimmNV sequence (KX883975.1, strain SCM51502), from Shi *et al.* 2016 (Fig. 4.2). Based on this phylogeny, comparisons to public sequences and to sequences from wild collected flies in Edinburgh, in this region of the genome there seems to be no detectable differentiation between the DimmNV infecting Dmel and Dimm.

I also observed transmission of DimmSV from wild Dimm into 6 Dmel OreR females (3.4% of exposed flies tested positive for the virus). I Sanger sequenced regions of the L (RdRp) and N (nucleocapsid) gene of DimmSV. Five DimmSV L gene sequences (630bp) from infected flies (4 donor vials, and 1 recipient) all were identical at the nucleotide level, and were 100% identical to KR822814.1 (DimmSV infected flies collected in the UK as part of Longdon *et al.* 2017) at the nucleotide level. This indicates that these infections do not represent divergent strains of DimmSV in comparison to previous studies. I built a phylogeny of these L gene sequences (Fig. 4.3), including 12 DimmSV L gene sequences from earlier chapters of this thesis, and two sequences from other studies (Longdon *et al.* 2011c; Shi *et al.* 2016b). Despite

there being no variable sites in the sequences from this chapter, and those from earlier chapters of this thesis, the phylogeny seems to place the DimmSV sequence infecting the Dmel OreR recipient closer to wild DimmSV infected Dmel from Scotland, rather than close to its Dimm Donor sequence. It also suggests that a single host shift might have spawned the wild and experimental Dmel infections. However, the low variability in the sequences, and lack of samples could create an inaccurate and partial picture of viral evolution, and indeed, this picture is refuted by the DimmSV N gene tree.

The 4 DimmSV N gene sequences (~525bp) from infected flies (4 recipients) were differentiated amongst each other by only 3 variable sites, all of which are synonymous. All sequences were >99.7% identical to at the nucleotide level to KX353094.1 (DimmSV isolate CMF33 N gene, collected in the UK as part of Longdon *et al.* 2017). I combined these sequences with 11 DimmSV N gene sequences from chapter 2, 18 DimmSV N gene sequences from Longdon *et al.* 2017, and a DimmSV sequence from Shi *et al.* 2016, to create a phylogeny of the DimmSV N gene across Dmel and Dimm. With more variable sites, and more sequences across the phylogeny, the Dmel OreR ImmSV recipient sequences seem to nest within the phylogeny of Dimm infecting sequences, rather than within a Dmel specific branch of the phylogeny. Overall, the Sanger sequencing data suggests that DimmNV and DimmSV RNA is present in wild Dmel individuals, and that they form part of the currently recognised diversity of these viruses, rather than new, Dmel specific clades.

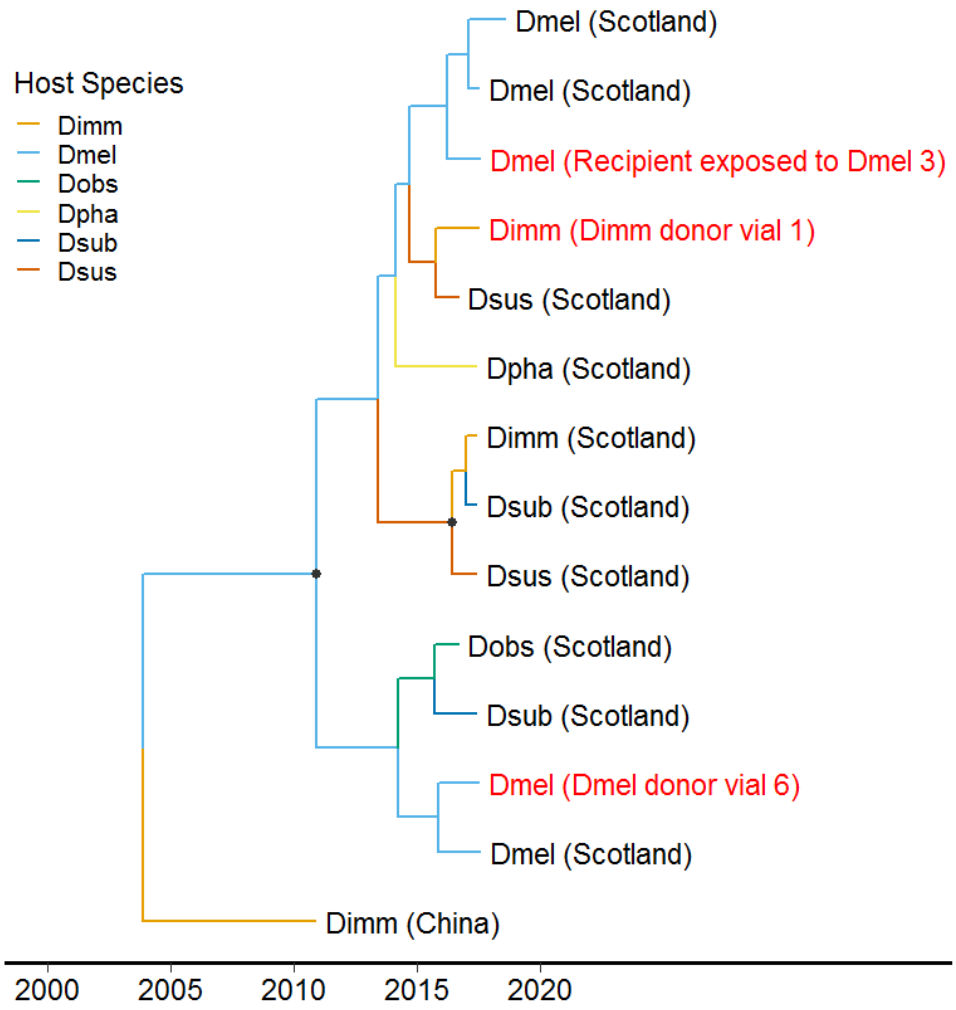


Fig. 4.2 **DimmNV maximum clade credibility tree including experimental and wild collected flies.** The tree shows the phylogenetic relationships between a ~600pb region of VP2 (the putative RdRp) of DimmNV. Host flies were from donor and recipient infections in the experiment, and wild-collected Drosophilidae from previous chapters of this thesis. Branches are coloured by host species, and node circles are present if posterior support >= 0.6. Tip labels display the host species and collection country. The three highlighted (red) tip labels are DimmNV infections from this experiment, 2 donor vials (1 Dmel and 1 Dimm), and one recipient OreR Dmel.

2265

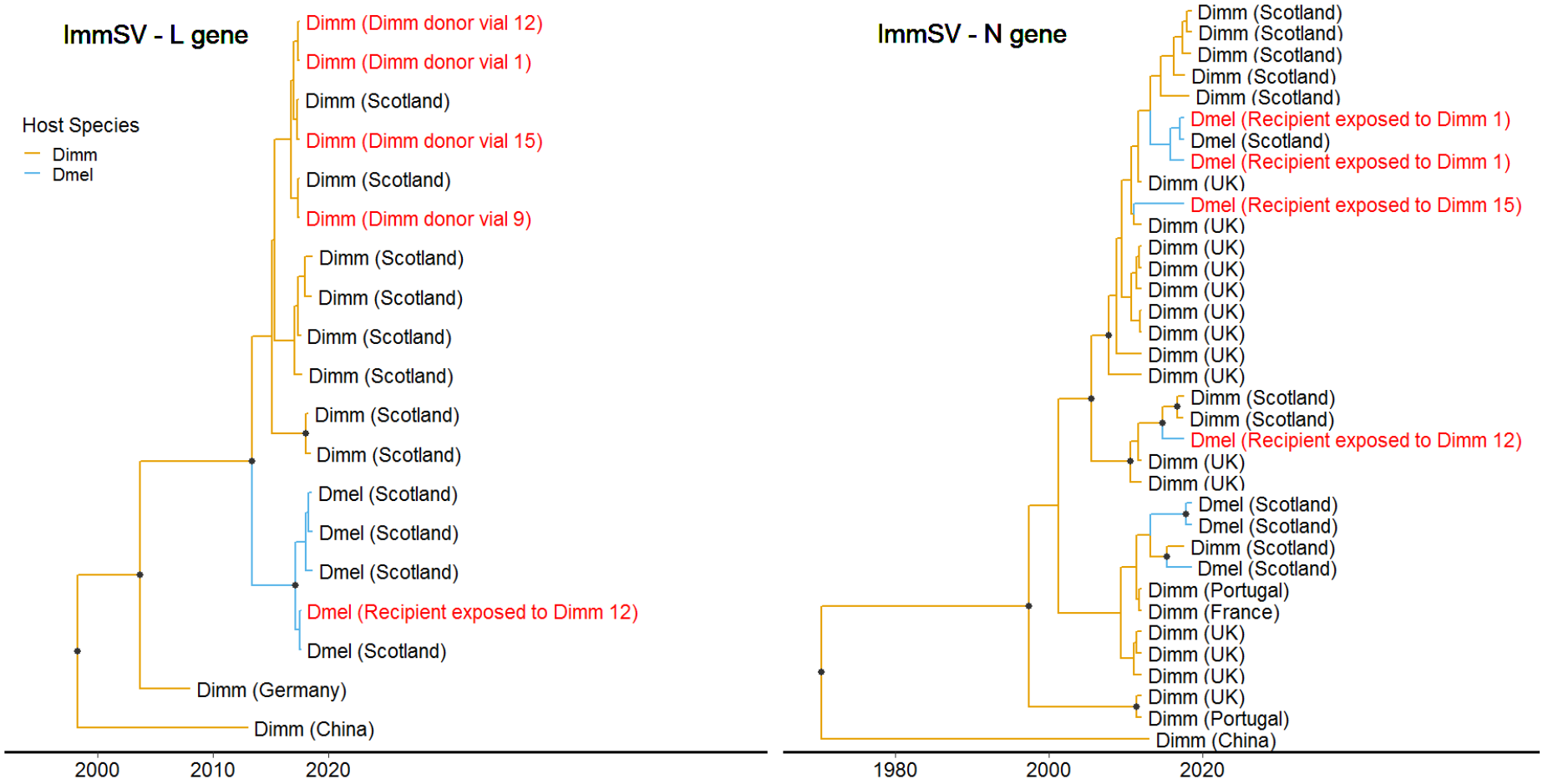


Fig. 4.3 Phylogenies of ImmSV sequences (regions of the L and N genes) from experimental and wild collected flies. The trees show the phylogenetic relationship between a ~625 bp region of the L (RdRp) and a ~525bp region of the N (nucleocapsid) genes of DimmSV. Host flies were from this experiment, wild-collected *Drosophilidae* from previous chapters of this thesis, and collections made as a part of other

studies (Longdon *et al.* 2011c; Shi *et al.* 2016b; Longdon *et al.* 2017). Branches are coloured by the most likely ancestral host species, node circles are present if posterior support >= 0.6, and tip labels display the host species and collection country. The highlighted (red) tip labels are DimmSV infections from this experiment, including donor vials (wild-collected flies), and experimental recipients (female Dmel OreR).

2266

4.3.2 Four naturally-occurring *Drosophila* viruses significantly reduce lifespan in *D. melanogaster*

I then examined the lifespan of the 606 virus-exposed Dmel females, some of which were infected with one or more of the nine transmitted viruses. Females showed faster declining mortality curves when infected with a virus verses than when uninfected. (Fig. 4.4).This suggests that viral infections could be reducing lifespan, without the need for transmission by systemic injection. To test this, I modelled the lifespan of Dmel females when infected by each of the transmitted viruses, using a linear mixed effects model. The model attempted to explain variance in lifespan by including all ten transmitted viruses as fixed effects, and exposure vial as a random effect. I compared the output and predictions from this model to that of a model which included type of virus as a random effect, and number of viral infections as a fixed effect. DIC values for the virus as fixed and random effects models were 5046.28 and 5040.17 respectively. Full output for both models can be found in Table 4.2.

Infection with four viruses significantly reduced lifespan in female Dmel: DmelNV, DimmNV, Muthill virus and Prestney burn virus. The estimated posterior distribution of each of these effects is displayed in Fig. 4.5, and posterior means and 95% credible intervals in Table 4.2. Of note, females that tested positive for DimmNV and Prestney burn virus, the viruses associated with the largest reduction in lifespan, had lifespans that were 29.8% (17.6% - 42.6%) and 36.4% (22.6% - 52.6%) shorter respectively. The variance associated with vial (which includes variance associated with exposure species) explained only 3.6% (95% HPD = 3.6×10^{-4} - 8.8%) of variance in lifespan, with a posterior distribution suggesting that this effect is not distinguishable from zero. As a summary of these results, Fig. 4.6 shows the lifespan of flies infected with each of the viruses, alongside the predictions from both models. In the random effects model, virus species caused explained only 5.54% (3.478×10^{-8} – 19.063) of variance in lifespan, around the effect of being infected by any virus, with a posterior distribution suggesting this effect is indistinguishable from zero. However, an increasing

number of viral infections was associated with a reduced lifespan (posterior mean = -7.5, 95% CI = -11.66 — -3.53) (Fig. 4.7).

For DmelNV, I additionally assayed a subset of the accompanying males and offspring, to test the impact of infections missed due to possible viral clearance within female lifetime. DIC for this model was 5050.677. I found that when data on infections in accompanying males and offspring was included in the analysis, it did not change the overall results (Fig. S4.3). However, the effect size of DmelNV was reduced (posterior mean = -4.885, 95% CI = -9.65, -0.45).

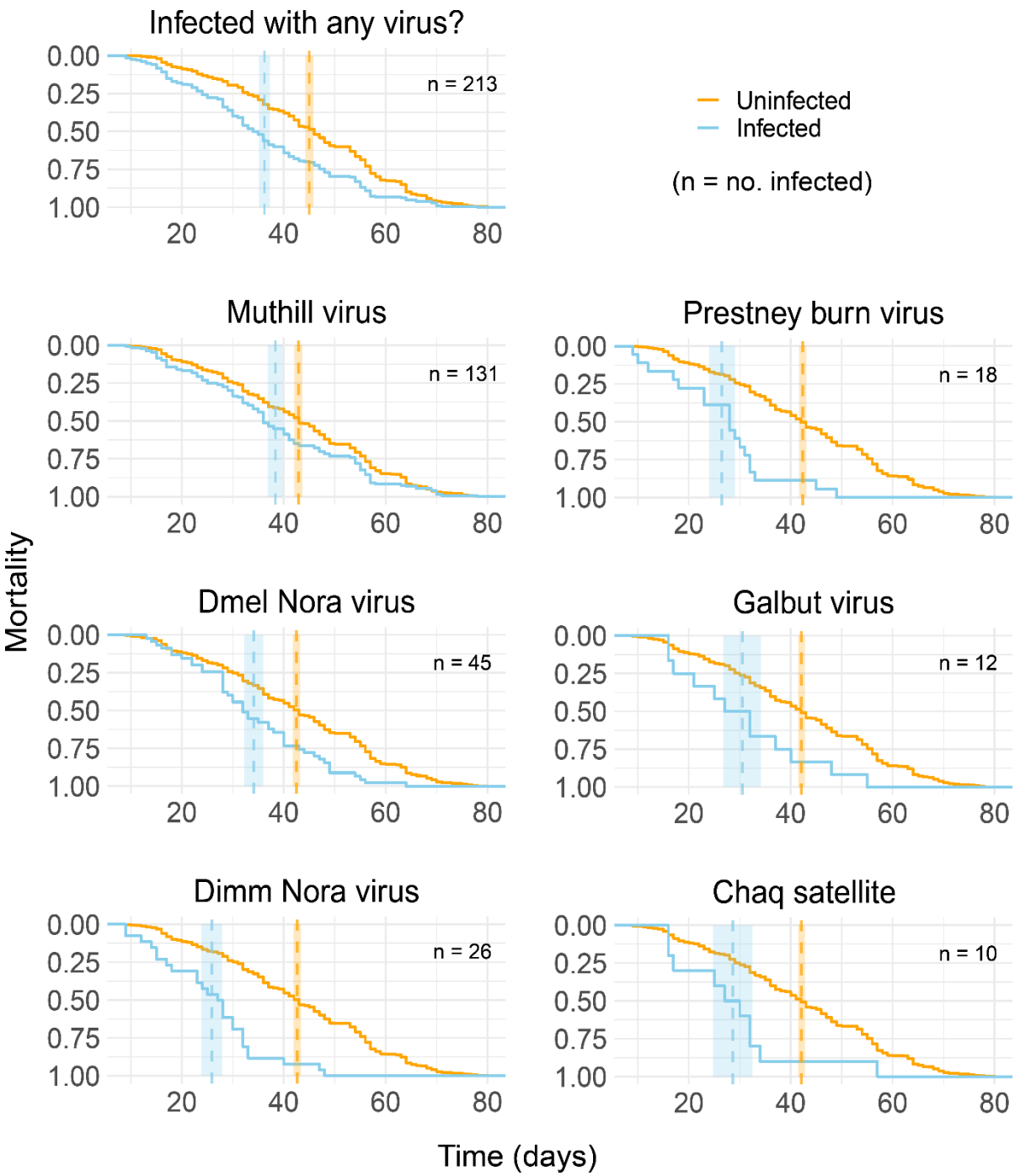


Fig. 4.4The effects of viral infection on mortality over time in Dmel OreR females. The plot panels show comparisons between the mortality curves of infected Dmel OreR females (blue line; exposed to viral infection through contact, and assayed by RT-PCR) and uninfected (orange line, the same uninfected flies appear on each graph). The n shown on each graph indicates the number of individuals used to estimate the blue lines (infected). The vertical dotted lines indicate the mean lifespan of a fly in each category, with shaded standard errors. Only viruses with at least 10 infections are included.

2304

Variable	Effect type	Posterior Mean	lower 95% Credible Interval	Upper 95% Credible Interval	MCMC Effective sample size	MCMC p-value
Model 1 (viruses as fixed effects)						
Intercept	**fixed**	**44.90**	**43.14**	**46.68**	**10304**	**<0.001***
Chaq satellite	fixed	-9.55	-21.02	2.42	10000	0.103
Galbut virus	fixed	-5.41	-15.85	5.57	10000	0.325
Muthill virus	**fixed**	**-4.90**	**-8.27**	**-1.36**	**10000**	**0.004***
***D.immigrans* Nora virus**	**fixed**	**-13.38**	**-19.99**	**-6.60**	**9730**	**<0.001***
***D.melanogaster* Nora virus**	**fixed**	**-7.81**	**-13.16**	**-2.79**	**10000**	**0.0028***
D.immigrans Sigma virus	fixed	-2.95	-16.79	10.00	10000	0.678
Grom virus	fixed	1.79	-21.09	24.40	10000	0.868
Prestney burn virus	**fixed**	**-16.35**	**-24.54**	**-8.62**	**9646**	**<0.001***
Corseley virus	fixed	13.08	-9.62	34.21	10226	0.241
Larkfield virus	fixed	-12.20	-35.55	11.10	10000	0.311
Experimental vial	random	8.69	0.00	21.90	6216	NA
Residual variance (units)	residual	232.51	206.02	261.51	9183	NA
Model 2 (viruses as random effects)						
Intercept	**fixed**	**45.06**	**43.30**	**46.96**	**10000**	**0.0001***
No. of viral infections	**fixed**	**-7.50**	**-11.66**	**-3.53**	**10000**	**0.0062***
Virus species	random	15.88	8.49×10^{-8}	56.48	9805	NA
Experimental vial	random	12.77	1.63×10^{-5}	26.00	10000	NA
Residual variance (units)	residual	230.43	204.27	258.07	10000	NA

Table 4.2 Outputs from linear mixed effects models used to analyse the effect of viral infection on lifespan in female Dmel. Fixed effect predictors were determined to be significant if their 95% credible interval did not overlap zero. Both models utilised a Gaussian distribution, and therefore posterior means can be interpreted on the data scale (days).

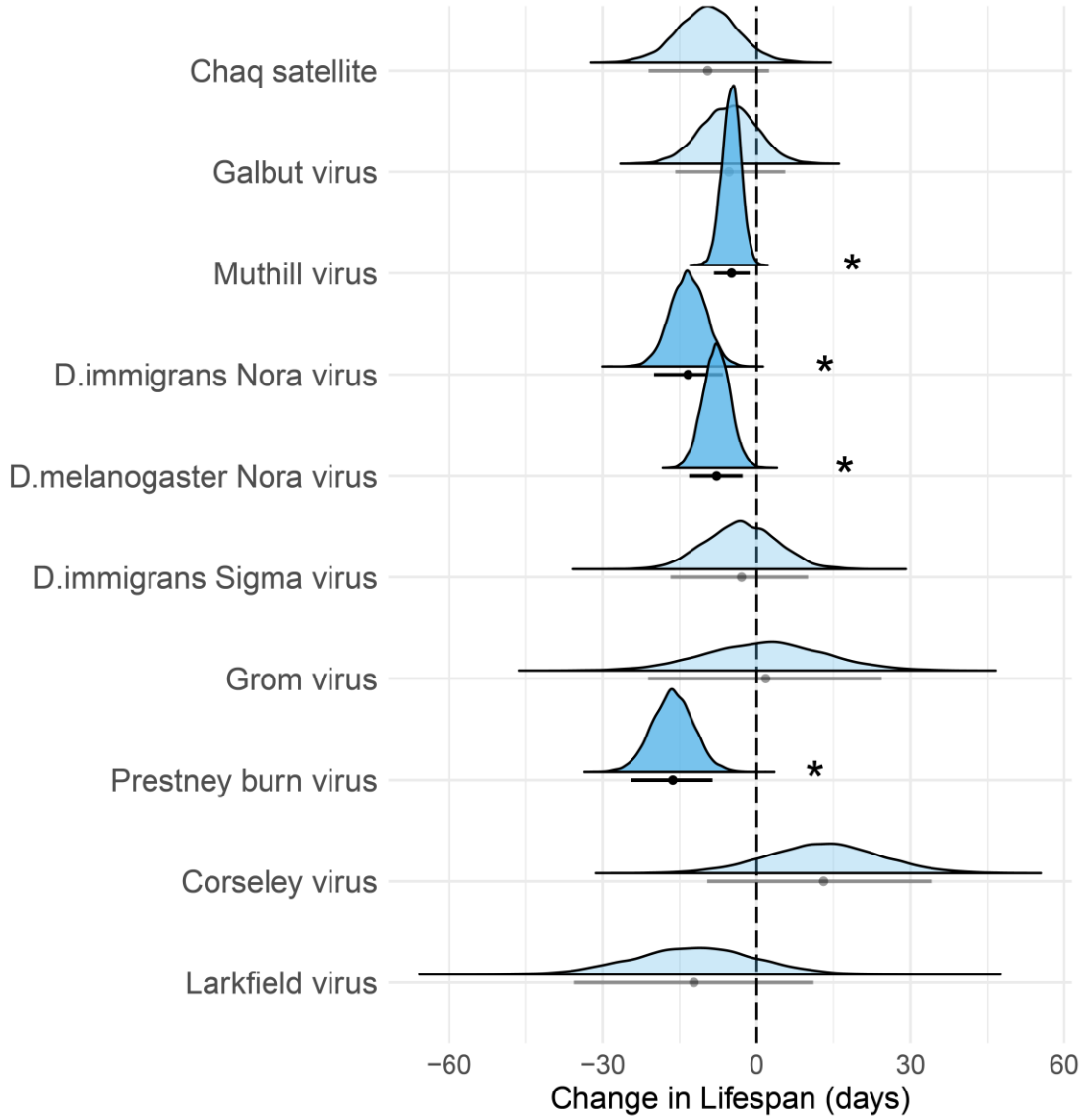


Fig. 4.5 Posterior density distributions of the viral fixed effects included in the model of lifespan variation in Dmel females. Solid horizontal lines and points indicate the 95% credible intervals, and posterior means, for each of the fixed effects included in the model. Above these estimated posterior density is displayed as output from MCMCglmm.

2308

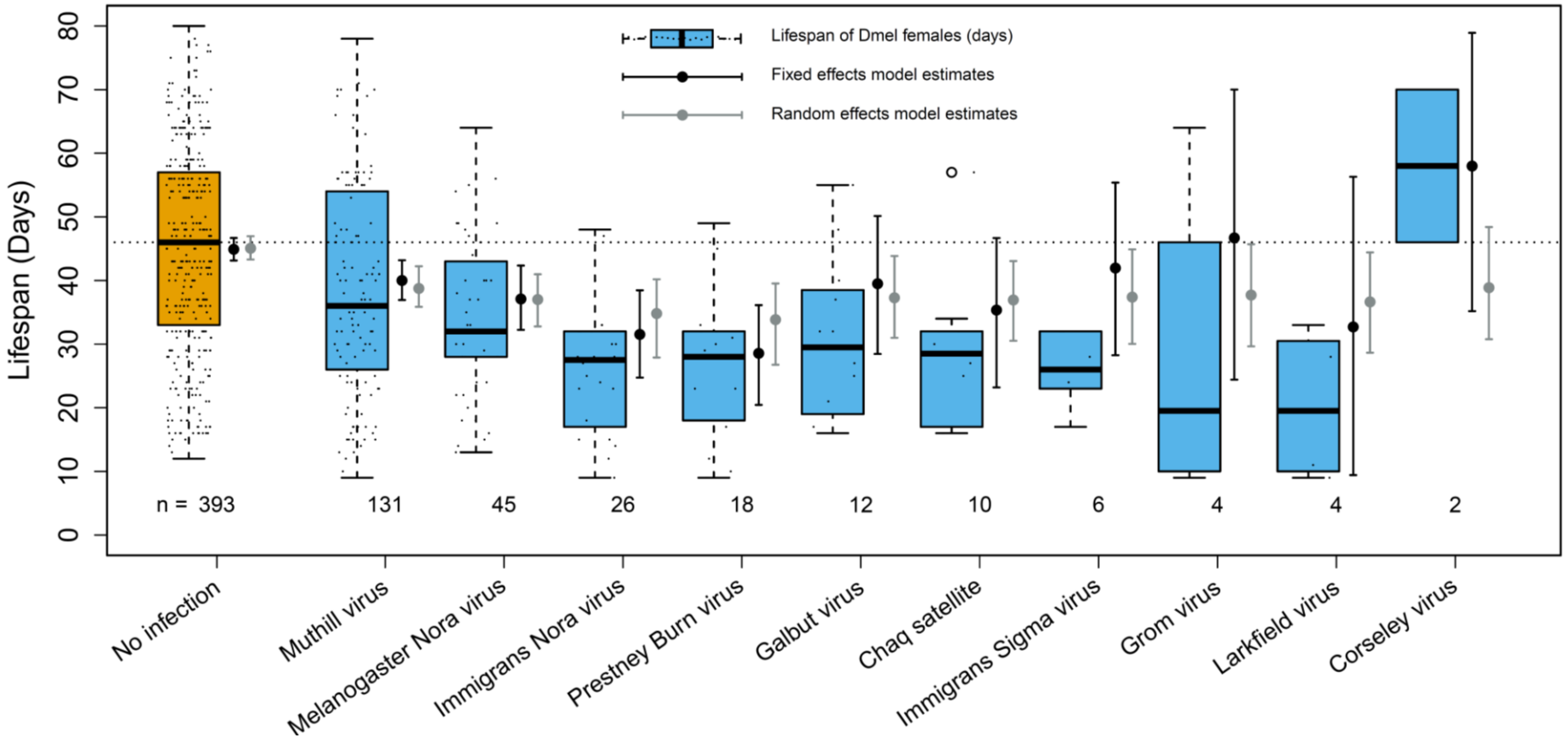


Fig. 4.6 The effect of viral infection on lifespan in female Dmel (OreR). The plot shows the lifespan of Dmel females when clear of viruses (orange), and when infected by ten different viruses (blue) as boxplots. The number of individuals infected with each virus is displayed under the boxes, and alongside are the predictions from the linear mixed effects models.

2309

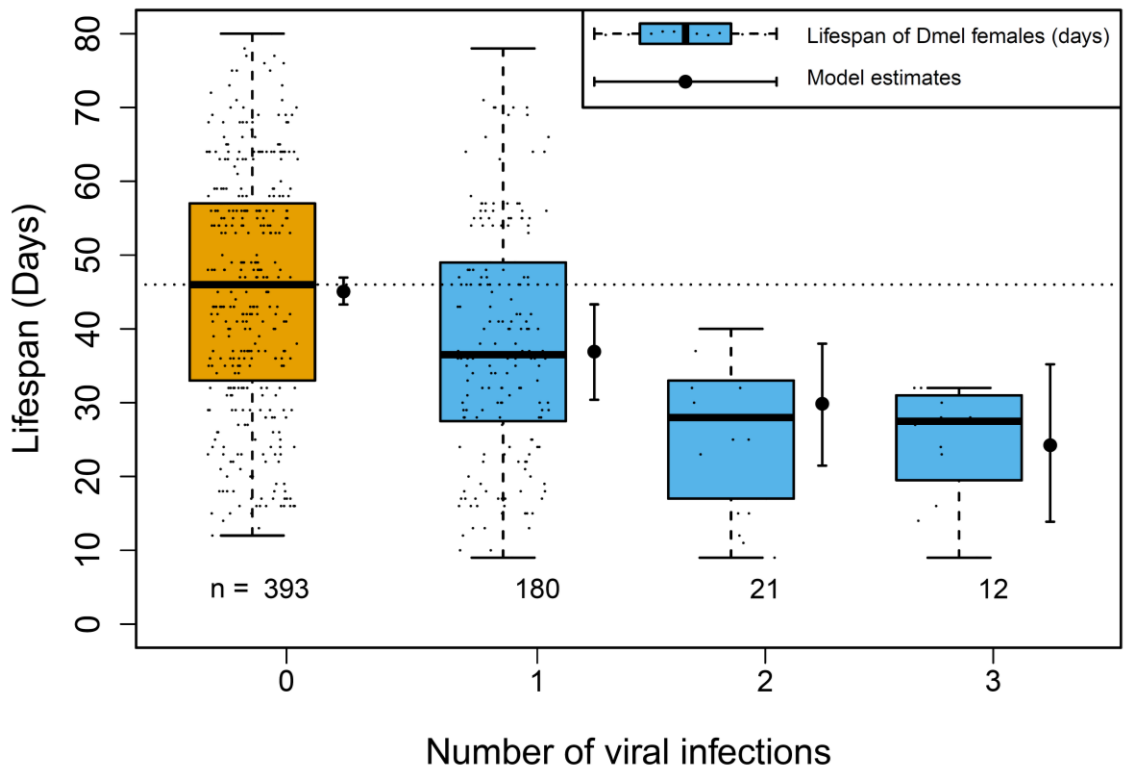


Fig. 4.7 The effect of the number of viral infections on lifespan in female Dmel. The plot shows the distribution of female lifespan observed when uninfected, or multiply infected with viruses. The number of Dmel females in each category is displayed below the boxplots. Model predictions come from the 'viruses as random effects' model (Model 2), in which number of viral infections was a significant predictor of lifespan.

4.3.3 Three naturally-occurring *Drosophila* viruses significantly reduce lifetime offspring production from *D. melanogaster* females

Selection acts on changes in total lifetime fecundity, which can be due to differences in lifespan, or to differences in the rate of egg production over the course of life. *Drosophila* viruses are known to affect both components of fitness (eg. Arnold *et al.* 2013b; Palmer *et al.* 2018), but they have rarely been studied together. It's is important to do so because trade-offs between them can easily be misinterpreted. For example, if early fecundity alone is followed, terminal investment could be misinterpreted as an increase in fitness, when it is really a

mitigation strategy. Therefore, I examined the offspring production associated with viral infections in this experiment.

The mean number of offspring produced by females infected by any virus was lower than that of an uninfected female (uninfected = 101.55, infected = 75.15), and offspring appeared to be produced at a slower rate (Fig. 4.8).To test this hypothesis, I used a hurdle-Poisson model to ask whether any specific viral infections are associated with a decrease in the total lifetime number of offspring produced by Dmel females. This model, which accounts for zero-inflation in the data, provides not only the effect of viral infection on non-zero offspring production, but also any change viral infection has on the likelihood of sterility. The DIC of the model was 4914.76. Again, the species or vial to which the Dmel females were exposed had a negligible effect on offspring production, explaining only 1.772% of variance (95% HPD interval: 0.019 - 5.328%), and with a posterior distribution piled up against zero. Infection with three viruses, DmelNV, DimmNV and Muthill virus, was significantly associated with a reduced number of offspring produced by Dmel females over their lifetime. Most strikingly, the model predicts a decrease in lifetime offspring production of 46.6% (36.4% - 57.9%) associated with DmelNV infection, the virus with the largest effect. I also found that infection with Muthill virus was associated with an increased likelihood of producing zero offspring. Regression coefficients, and p-values for the hurdle-Poisson model can be found in Table 4.3, and posterior distributions of the viral fixed effects in Fig. 4.9 As a summary of these results, Fig. 4.10 shows the lifetime offspring production of flies infected with each of the viruses, alongside the predictions from the model.

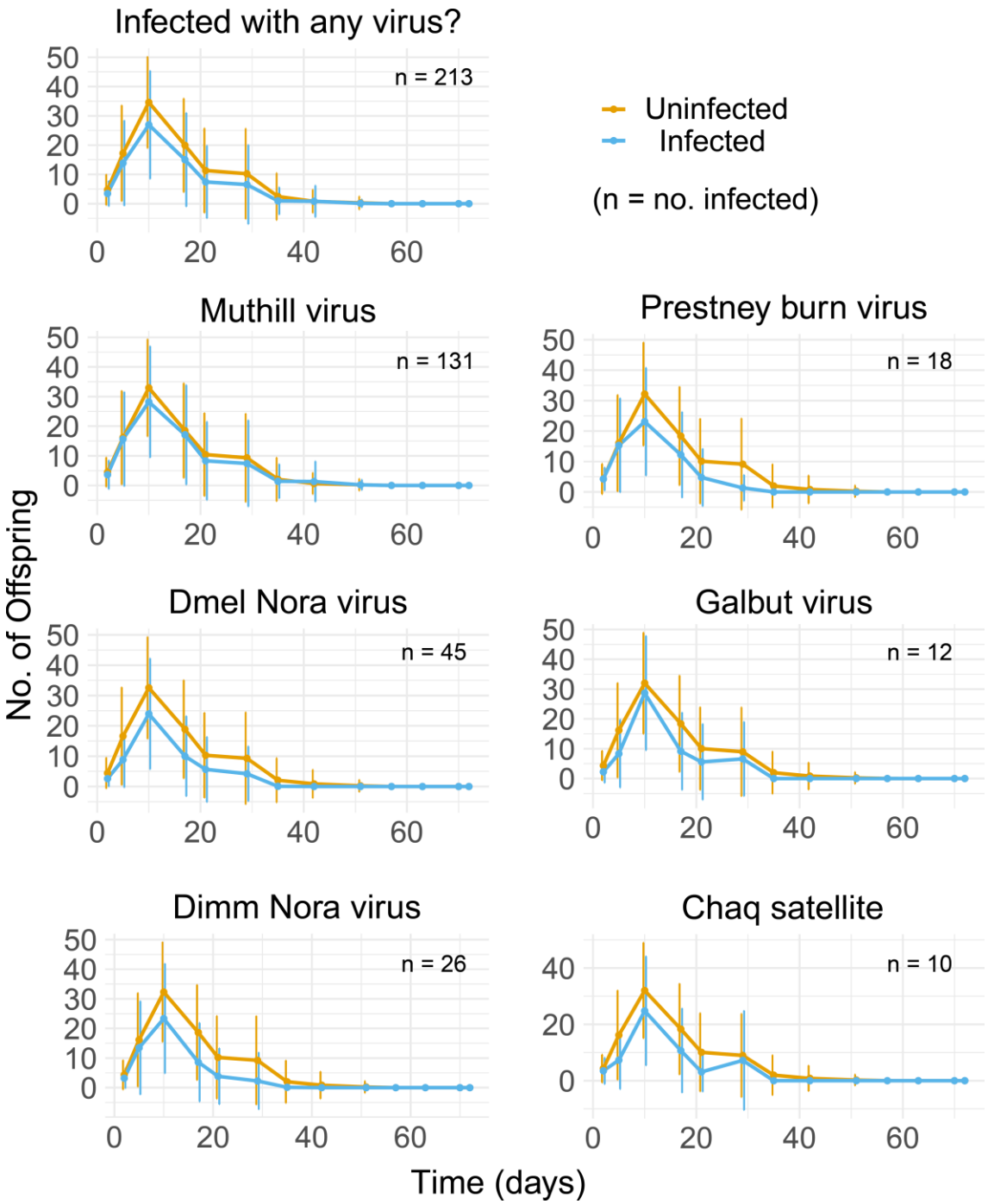


Fig. 4.8 No. of offspring produced by Dmel females, over time, when infected with viruses. The plot panels show comparisons between the offspring produced over time by uninfected and infected Dmel females (exposed to viral infection through contact). The curves are coloured by infection status, points indicate the mean, and bars indicate the standard deviation around the mean, per day of offspring collection. The top left plot shows a comparison between flies infected with any virus, and those clear of all viruses they were exposed to. All other plots show a comparison between the flies infected with one particular virus, and the rest of the flies in the experiment. Only viruses with at least 10 infections are included.

This reduction in total offspring number might simply be a consequence of reduced lifespan. Alternatively, it may be due to a lower rate of egg-laying, even in early life. In the wild, this loss of early reproduction may have a disproportionate effect, as few flies are likely to live beyond 2 weeks. Therefore, I also examined the effect of viral infection on the number of adult offspring produced from the first 5 days of the experiment (females aged 13 days old). I excluded any females that died before this date, so that lifespan had no effect on this analysis, but used otherwise the same model as the lifetime offspring production analysis. The DIC of the model was 3869.512. Even in this analysis, where the impact of longevity is excluded, DmelNV was still significantly associated with a reduction in offspring production (Poisson regression coefficient = -0.36, 95% CI = [-0.685, -0.024], pMCMC = 0.032 *). See Fig. S4.4 for a plot of the posterior distribution of viral fixed effects in this model.

Variable	Effect type	Posterior Mean	lower 95% Credible Interval	Upper 95% Credible Interval	Effective sample size	MCMC p-value
Zero-truncated Poisson regression coefficients						
Intercept	**fixed**	**4.116**	**3.359**	**4.862**	**153873.36**	**< 0.0001 ***
Chaq satellite	fixed	-0.273	-1.026	0.479	153161.83	0.4765
Galbut virus	fixed	-0.069	-0.750	0.602	165344.32	0.8433
Muthill virus	**fixed**	**-0.242**	**-0.455**	**-0.031**	**167132.50**	**0.02494 ***
D.immigrans Nora virus	**fixed**	**-0.468**	**-0.936**	**-0.005**	**125058.17**	**0.04943 ***
D.melanogaster Nora virus	**fixed**	**-0.620**	**-0.938**	**-0.299**	**171532.33**	**0.00024 ***
D.immigrans Sigma virus	fixed	0.328	-0.529	1.193	176921.16	0.4553
Grom virus	fixed	-0.245	-1.704	1.225	161424.81	0.7425
Prestney burn virus	fixed	-0.450	-0.971	0.060	162432.57	0.0868
Larkfield virus	fixed	-0.729	-2.232	0.806	166417.24	0.3445
Corseley virus	fixed	0.763	-0.552	2.095	200000.00	0.2598
Experimental vial	random	0.016	0.000	0.048	45653.15	NA
Residual variance (units)	residual	0.875	0.760	0.996	147542.54	NA
Probability of zero offspring (logit scale)						
Intercept (hurdle)	fixed	-2.749	-5.679	0.117	2781.74	0.0464
Chaq satellite	fixed	1.621	-1.346	4.435	2842.10	0.2753
Galbut virus	fixed	0.838	-2.196	3.647	2434.47	0.5533
Muthill virus	**fixed**	**1.177**	**0.139**	**2.216**	**1504.19**	**0.0279 ***
D.immigrans Nora virus	fixed	1.232	-0.218	2.628	4201.37	0.1010
D.melanogaster Nora virus	fixed	0.455	-1.196	2.003	1859.71	0.5400
Grom virus	fixed	1.557	-4.172	7.016	3008.02	0.5544
Prestney burn virus	fixed	1.178	-1.204	3.354	2303.41	0.3022
Larkfield virus	fixed	0.328	-5.660	6.134	3113.88	0.8963
Experimental vial	random	< 0.001	< 0.001	< 0.001	0.00	NA
Residual variance (units)	residual	1.000	1.000	1.000	0.00	NA

Table 4.3: Outputs from the hurdle-Poisson mixed effects model used to analyse the effect of viral infection on lifetime offspring production in female Dmel. The probability of zero offspring was only allowed to vary with infection for eight viruses, as no zero-offspring individuals were infected with DimmSV or Corseley virus.

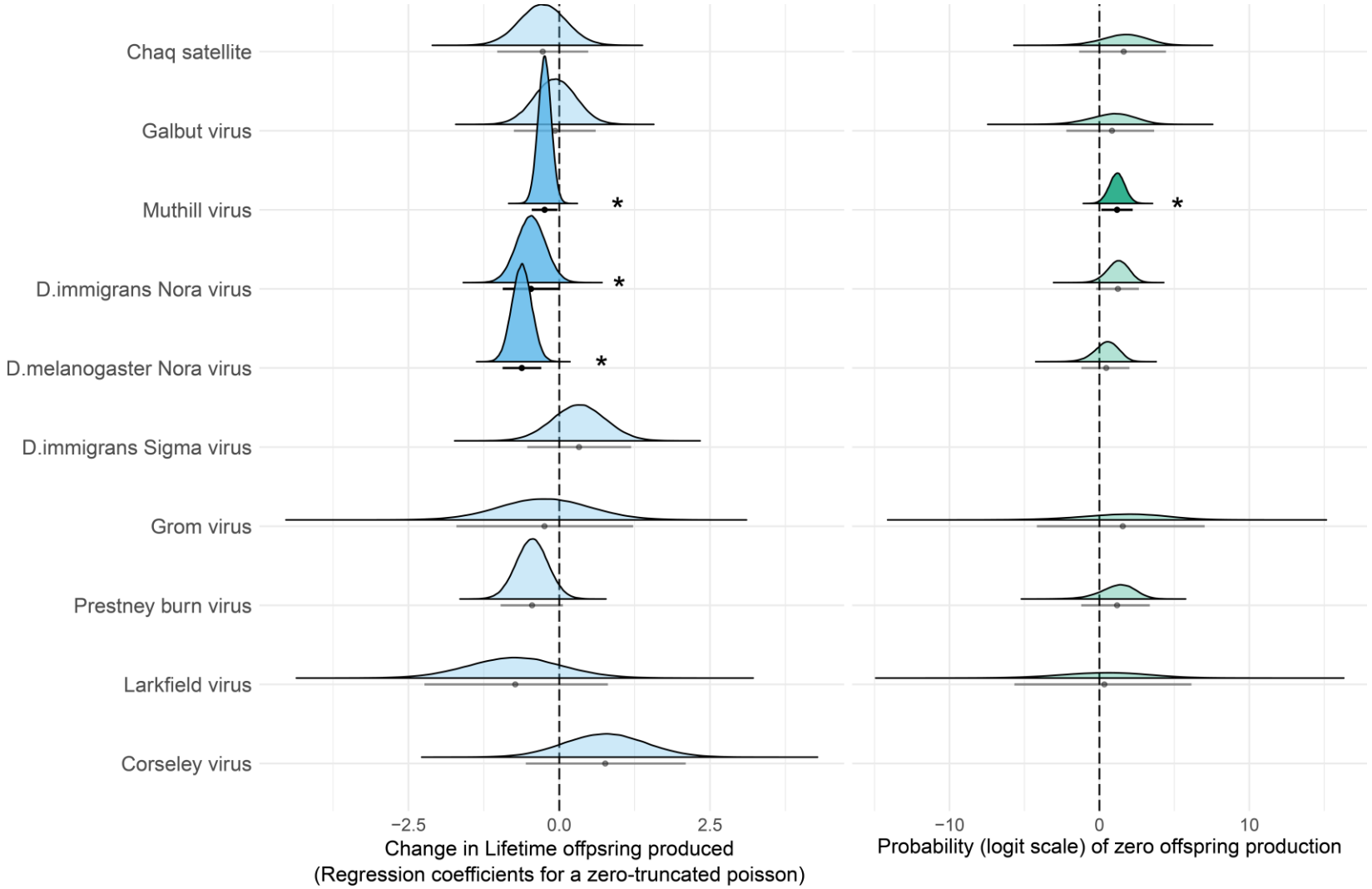


Fig. 4.9 Posterior density distributions of the viral fixed effects included in the hurdle-Poisson model of lifetime offspring production in Dmel females. Solid horizontal lines and points indicate the 95% credible intervals, and posterior means, for each of the fixed effects included in the model.' Above these the distribution of the estimates is displayed as output from MCMCglmm.

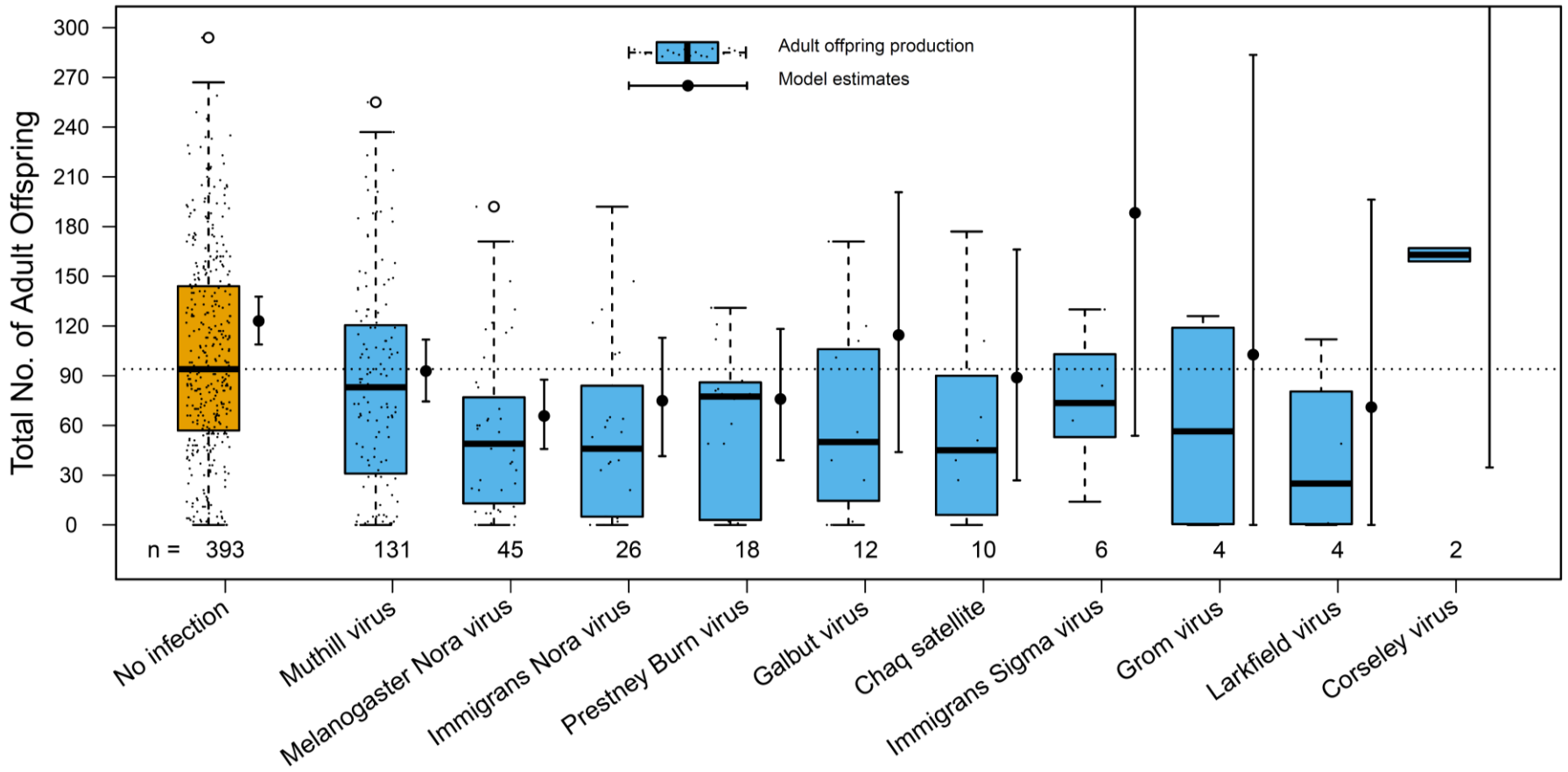


Fig. 4.10 The impact of viral infection on lifetime offspring production in female Dmel (OreR). The plot shows the number of adult offspring produced by Dmel females when clear of viruses (orange), and when infected by one of ten different viruses (blue) as boxplots. The number of individuals infected with each virus is displayed under the boxes, and alongside are the predictions from the linear mixed model (a hurdle-Poisson model).

2351

4.4 Discussion

4.4.1 Cross-species transmission of Drosophila viruses

Most experiments that characterise the pathology and life history effects of *Drosophila* viruses use systemic injection to infect flies. This not only limits experiments to viruses that have been isolated, but may also cause different infection outcomes than natural transmission routes (see review Mondotte & Saleh 2018). In this experiment, I attempted to bypass the need for systemic injection, infecting Dmel females with viruses through contact alone.

Assays using RT-PCR showed that the wild *Drosophila* collected as virus 'donors' contained 23 different viruses, of which 21 were found in only one species group. Host species and prevalence were broadly consistent with earlier surveys (Webster *et al.* 2016 and chapter 2; Table 4.1). The transmission of nine of these viruses into Dmel, seven across species, suggests that some *Drosophila*-associated viruses have the ability to infect multiple, distantly related species through contact alone. Divergence from the *D.melanogaster* subgroups of *D.immigrans* is estimated at 33.1± 3.16 MYA, and *D.obscura* at 24.9 ± 2.88 MYA, (Russo *et al.* 1995). However, in this experiment, the co-housing of multiple species in a densely populated vial, where prior viral shedding would amplify exposure, probably increased the likelihood of transmission events. This data concurs with previous data from chapter 2, which suggests that *Drosophila* viruses differ in their level of specialism to specific hosts. In general, viruses from the *Obscura* group were less likely to be transmitted to Dmel than Dimm or Dmel hosted viruses, concurring with observations *D.obscura* group are less likely to be shared with more distantly related hosts (Webster *et al.* 2016). However, Prestney burn, Corseley and Grom virus, which were also present in Dmel reads in Webster *et al.* (2015), bucked this trend and transmitted to Dmel females. Based on this data, further studies could look for evidence of the replication of these viruses in Dmel, using siRNAs as in Webster *et al.* (2015), and investigate whether viral titre is reduced with phylogenetic distance from the original host species, as in Longdon *et al.* (2011).

Two viruses showed a particular propensity to infect other host species, DimmNV and Muthill virus, infecting all three host species groups in the wild, and transmitting to Dmel from Dimm and Dmel in the lab. DmelNV, the close relative of DimmNV, infects the midgut cells of infected flies (Ekström & Hultmark 2016), and is transmitted via the faeces (Habayeb et al. 2009a). Therefore, it is likely that DimmNV transmission also occurred faecal-orally. However, a previous study found that DimmNV viral protein 1, which contains a viral suppressor of RNAi, is unable to supress the cleavage activity of Dmel Argonaute-2 (Ago-2) in Dmel S2 cells, in contrast to DmelNV, which has this ability (van Mierlo *et al.* 2014). This seems to be due to its inability to interact with the Dmel Ago-2, and suggests that in Dmel, DimmNV would be suppressed, or at least repressed, by the RNAi response. However, in this experiment, the observation of DimmNV infection in wild Dmel donors (prevalence = 9.3%, [6.1, 29.4%]), and transmission from Dimm to Dmel in the lab (18.37% of exposed Dmel), suggests that DimmNV has the ability to overcome this suppression, and establish itself in Dmel (see phylogeny of DimmNV infections Fig. 4.2). Indeed, these observations agree with data from chapter 2 of this thesis, where I found DimmNV at a global prevalence of ~ 30% in wild Dmel populations around Edinburgh.

I observed evidence for horizontal transmission of two viruses previously thought to be obligately vertically transmitted: Galbut virus, and DimmSV. DimmSV RNA was detectable in 6 Dmel females exposed to Dimm (3.4% of those exposed). Previous studies have reported this virus to be obligately vertically transmitted, with particularly high rates of maternal transmission (Longdon *et al.* 2017). Galbut virus was also reported to be primarily vertically transmitted by Cross *et al.* (2020). This study found insufficient evidence for Galbut virus replication after ingestion of an infectious homogenate as, in three flies, the levels of virus RNA present 21 days after ingestion was not significantly different from levels present immediately after ingestion. However, in this experiment I detected Galbut virus RNA, and RNA from its Chaq satellite or optional segment, in 12 and 10 of the exposed Dmel females

respectively (transmission to 8.82%, and 7.35% of exposed flies). These observations could be due to genuine, horizontal transmission of these viruses between host species, or alternatively, the presence of residual, non-replicating virus RNA from ingestion. I did not attempt to quantify viral replication in the Dmel OreR females, and therefore it is possible that I was simply detecting residual, non-replicating virus RNA from ingestion. Nevertheless, flies were only exposed for three days, and RNA remained at the time of death (mean 22.5 and 19 days later, in the case of Galbut, and DimmSV respectively). This makes it seem unlikely that viral RNA on the outer body of flies, or ingested but not replicating RNA, wouldn't be excreted or shaken off by tipping in this time period. Additionally, the DimmSV infections in Dmel observed in this experiment, and confirmed with Sanger sequencing (Fig. 4.3), concur with data from earlier chapters of this thesis, which found a ~4% prevalence of DimmSV in wild Dmel populations around Edinburgh. Further experiments to confirm the replication of DimmSV in Dmel tissues will be needed to support these observations, and in particular, infection of non-reproductive tissues, which might enable horizontal transmission.

4.4.2 Infection with four viruses reduces lifespan in D.melanogaster

Using a linear mixed effects model approach, I found that infection with four viruses (Muthill virus, DimmNV, DmelNV and Prestney burn virus), was significantly associated with decreased lifespan in Dmel females. All four viral infections reduced lifespan by >10%, and two by >20% (Muthill - 10.9%, DmelNV - 17.4%, DimmNV - 29.8%, and Prestney burn - 36.4%). No viral isolate exists for three of these viruses - Muthill virus (+ssRNA, cf. Negevirus), Prestney burn virus (+ssRNA, cf. Polerovirus/ Sobemovirus), and DimmNV (+ssRNA, Picorna-like virus) and this study presents the first evidence that these three viruses might reduce lifespan in Dmel females. Of note, DimmNV VP1 is unable to suppress the Dmel RNAi response (see above), therefore it's possible that it might show higher titres in Dimm than Dmel. However, qPCR of DimmNV infected Dmel will be needed to determine if this is the

case. It's possible that other viruses were also associated with reduced lifespan, but with the number of transmission events, I did not have enough power to detect them.

In comparison to the other costly viruses, DmelNV, a +ssRNA picorna-like virus, has been isolated, and was previously thought to have no pathological effects. Habayeb et al. (2009a) found that DmelNV infected OreR survival was similar to that of uninfected flies up to 50 days. Survival after this point seemed to decrease in infected flies faster than in uninfected flies, but the significance of this difference was uncertain as measurements were halted before all flies were dead. A later study found that DmelNV infected flies do show differential expression of Toll and IMD pathway associated genes (Cordes et al. 2013) suggesting that flies are having to devote energy into mounting an anti-viral immune response. So perhaps higher-powered experiments, and lifetime observation of infected flies is needed to detect fitness effects of DmelNV. An increase in the number of viral infections in Dmel females also reduced lifespan in a linear manner. However, larger sample sizes, and quantification of viral titres during co-infections will be needed to determine whether prior infections facilitate, inhibit, or accommodate co-infection in *Drosophila* hosts.

4.4.3 Infection with three viruses reduces lifetime offspring production in D.melanogaster

Using a generalised linear mixed model, I found that infection with three viruses (Muthill virus, DimmNV and DmelNV), was associated with a significant decrease in lifetime offspring production in Dmel females. I also found that infection with Muthill virus increases the likelihood of no adult offspring being produced. However, whether these effects happen because of lower egg production, or larval death is unclear. Habayeb et al. (2009) found no effect of DmelNV infection on either the number of eggs laid, or the eclosion rate (% adults hatching from eggs) of female Oregon R in infected and uninfected stocks. However, their study of eclosion rate used sample sizes of only 30 flies in each experimental group, and

perhaps, again, larger-scale experiments are needed to detect viral effects on fitness-related traits. A larger scale study of the eclosion rate of the offspring of infected and uninfected females would be needed to investigate this effect further. My statistical power to detect the effect of other viral infections on the likelihood of sterility, or offspring production, was low, and so these may not be the only viruses associated with a reduction in fecundity associated traits. However, for DmelNV, DimmNV and Muthill virus, these results present the first evidence that these viruses significantly decrease offspring production in Dmel. They also demonstrate that initial characterisation of fitness-related costs of viral infection is possible for non-isolated viruses.

Because of the shortened lifespan of wild flies (Rosewell & Shorrocks 1987), and to exclude the influence of lifespan, I also tested whether virus infection could reduce the offspring produced by Dmel females in early life (to day 5 of the experiment, all females aged 13 days). In contrast to the main offspring analysis, the only virus that was significantly associated with reduced offspring production in early life was DmelNV (posterior mean effect size = -0.36, 95% HPD [-0.685, -0.024]) (Fig. S4.4).This suggests that DmelNV infection leads to not only a reduced lifespan, but also a reduced rate of offspring production, even before death. It's probable that in the wild, predation, starvation, or death by other natural means kills flies before the end of their lifespan in the lab, and therefore, this viral-induced lifespan reduction wouldn't be so stark. However, without data on the lifespan of wild *Drosophila*, it's difficult to draw realistic conclusions. In contrast, there was no detectable effect of Muthill virus or DimmNV infection on early offspring production, most likely due to the lack of statistical power.

There are some examples of insect viruses with a commensal, or even mutualistic role in their host species, such as Rift valley fever virus in *Aedes aegypti* (Rossignol *et al.* 1985), and in *Drosophila,* DCV may induce flies to increase ovariole number (Thomas-Orillard 1984). Indeed, many insect viruses could be commensal in their hosts. However, it is not possible to conclude that the viruses which showed no negative effect on fitness components were commensal, or mutualistic, in Dmel. It's more likely that feasible experiments lack the power

to detect a negative (or positive) effect on fitness for most viruses. Viruses with <20% reduction in lifespan require large sample sizes (100-150 flies per treatment group) to observe significant reductions in fitness, even in a balanced experimental design (Fig S4.5). In fact, based on simulations using the mean and variance from this experiment, a sample size of >45,000 flies per treatment group would be needed to detect a 1% reduction in lifespan, which could still drive evolution in natural populations. These simulations demonstrate that to characterise the fitness cost of smaller-effect viruses, possibly the status quo for the insect virosphere, we will need to find a way to create large, high-powered experiments that still employ natural methods of virus transmission.

4.4.4 Caveats

There might be two sources of inaccuracy in my ability to assess the effect of viruses on Dmel lifespan, which would act in opposing directions on estimates. Both lie in the possibility that Dmel might be able to clear viral infections throughout their lifetime, as demonstrated in flies orally infected with DCV (Mondotte et al. 2018). First, if a fly simply dies earlier because it's a low quality fly, rather than it's lifespan being shortened by viral infection, and there’s a higher chance of viral detection if that fly dies earlier (because the virus is cleared by the fly), I might erroneously detect an effect of viral infection on lifespan which does not really exist (Fig S4.1, illustrating the clearance issue). Alternatively, if viruses are cleared by the flies, and then the fly dies, we wouldn't record a fly as infected, when in fact its lifespan may have been shortened by the virus. This would lead us to underestimate the viral effect on lifespan.

Habayeb et al. (2009b) found that when individual flies infected with DmelNV were followed for 19 days, the levels of viral RNA secreted into faeces either remained high across the timescale, or dropped dramatically over the first 4 days and then stayed low. If internal and external virus RNA levels are correlated, this suggest that a pattern of either titre-dependent persistence, or virus clearance exists. I therefore assessed the effect of sampling biases on

my lifespan data collection potentially induced by clearance of the viruses throughout lifespan, as demonstrated in flies orally infected with DCV (Mondotte et al. 2018). I included 16 additional infections in accompanying males and offspring in the counts of DmelNV presence/absence, and found that, although it decreased the posterior effect size of the DmelNV effect, infection still significantly reduced lifespan (Fig S4.3). This suggests that I did miss some virus infections in the RT-PCR assays, and that the likelihood of viral detection decreases throughout lifespan. However, whether this is because of viral clearance remains to be seen, as there currently is no other evidence for DmelNV, and no evidence at all for all other viruses suggesting clearance of infections. Overall, the inclusion of this data did not affect my conclusion that oral infection with DmelNV significantly decreases lifespan.

4.5 Conclusions

In this chapter, I found evidence that some *Drosophila* viruses may be extremely host generalist, with the ability to transmit across species in the lab through, most likely, faecal-oral transmission. This includes two species which were previously thought to be vertically transmitted, and further experiments will be needed to investigate if replication is happening in their new host species. I found that four of these infections significantly reduce Dmel female lifespan, and three, offspring production. Increases in the number of viruses infecting Dmel females also decreased lifespan. It seems that even viruses which are fairly common in wild Drosophila can be maintained when they exact a strong fitness cost on their hosts. Further studies could focus on the mechanism by which these reductions in lifespan and offspring production occur, whether they are as strong in outbred, or wild-collected flies, and whether they are driving any kind of selection in the host. It's possible that I have only detected the viruses with the strongest effects, and that the majority of infection phenotypes in the *Drosophila* virosphere are less extreme. However, characterising these subtler effects will require even larger, high-powered experiments, and probably oral infection methods which rely less on chance to determine sample size. But, with these experiments, we could start to describe the realistic, population level fitness costs of viral infections in insect hosts.

5 General Discussion

In this thesis, I set out to demonstrate the utility of the *Drosophila* and their naturally-occurring viruses for examining insect virus co-evolution in its natural habitat, within the complexity of multi-host, multi-virus systems (reviewed in other systems in Hall *et al.* 2020). By quantifying and characterising this complexity, we can gain meaningful context to the patterns of selection seen on insect, and insect virus genomes. Additionally, we can properly parameterise models of virus host range, abundance, and prevalence in insects. In this chapter, I will summarise my key findings. I will also outline some future directions for this research, which utilise this system, or the data collected from this thesis.

5.1 Discovery of Drosophila associated viruses

In chapter two, I use metagenomic sequencing of Scottish drosophilids to identify 17 novel viruses, and extend the genomes of 8 known viruses, expanding the known *Drosophila* virosphere by ~10%. Some of these viruses are of particular interest, such as Hermitage virus (Webster *et al.* 2016 - extended in Ch2), which is basal to the *Flaviridae* (Shi *et al.* 2016c), a family containing several human pathogens. However, the purpose of this was not only to expand the known *Drosophila* virosphere. It was to demonstrate the analyses and insights possible when the totality of viruses infecting a sympatric group of related host species is characterised. We will need to employ methods of virus discovery that rely less on similarity searches, possibly by using virus-derived siRNAs to discover 'viral dark matter' (eg. Obbard *et al.* 2020), to be confident that we've captured the full diversity of viruses infecting *Drosophila*. However, with the viruses already described, there are many exciting questions we can ask about insect-virus evolution. For example, several groups of *Drosophila* viruses that are each-others closest relatives, such as Drosophila Midmar tombusvirus (described in Ch2), and Dansoman virus (Webster *et al.* (2015), closely related to chronic bee paralysis virus), could be used for future studies of co-divergence, and co-evolution.

5.2 The host range and prevalence of Drosophila viruses

In chapter two I also examined the host range, and prevalence of 41 new and known *Drosophila* viruses. Highlighting the importance of examining co-evolution in multi-host, multi-virus systems, I found that 90% of viruses infect multiple host species, and that all host species with at least 4 individuals sampled host multiple viruses. This suggests that, within this system, most viruses are host generalists, though I was unable to distinguish between true, and 'apparently' multi-host pathogens (Fenton & Pedersen 2005). I also found that an appreciable proportion of individuals (>25%) were co-infected with multiple viruses, suggesting that co-infections are not a rarity in wild insect virus systems (also see Batson *et al.* 2020). This suggests that evolution of viruses to better infect, and transmit, among insects often involves them evolving to co-infect those individuals. Examining insect-virus co-evolution using solely single virus infections therefore does not capture the totality of virus-driven selective pressure on insect genomes, or indeed the selective pressure on virus genomes possibly driven by other, co-infecting, viruses.

I also found that virus prevalence, sampled for ten viruses, can vary significantly with host species. These patterns could be caused by the different susceptibilities of *Drosophila* species to infection with the same virus (eg. van Mierlo *et al.* 2014). However, there are far more explicit questions we can ask of this dataset, with more complex models and statistical methods. For example, previous studies in mammals have found host phylogeny to be a significant predictor of viral sharing (Davies & Pedersen 2008), and the frequency and success of viral host shifts (Faria *et al.* 2013). Additionally, in *Drosophila,* host phylogeny is a significant predictor of the susceptibility of host species to some viral infections (Longdon *et al.* 2011a). Viral phylogeny may also influence the titre which viruses can reach on infection in a new host, with closely related *Drosophila* viruses displaying more correlated titres (Imrie *et al.* 2021). However, evidence for this is based on testing with only four virus strains. The dataset I

collected in chapter two, which contains 10 viruses for viral presence/absence, and 41 for host range respectively, expands the data available for these kinds of analyses significantly. With it, we could use a co-phylogenetic mixed model (Hadfield *et al.* 2014) to ask whether there are linear, or non-linear, viral or host phylogenetic effects on host range or viral prevalence. I.e. do we see more similar patterns of viral prevalence, or viral presence/absence closely related hosts, or closely related viruses?

In chapter two I also found that, for some viruses, prevalence varies by collecting season. Studies of hytrosaviruses of tsetse flies suggest that insect viruses can also vary significantly in prevalence across collecting sites (reviewed in Kariithi *et al.* 2017), though initial analyses suggests this was not detectable in my data. To quantify spatial and temporal variation in virus prevalence in more detail, I plan to employ similar models as I used in chapter three to quantify spatiotemporal autocorrelation in virus prevalence (also see Myer & Johnston 2019), and ask whether virus prevalence varies significantly in space and time. Climatic and environmental factors that drive seasonal fluctuations in host abundance may also drive variation in virus prevalence. Therefore, if possible, it would also be interesting to investigate the role of temperature, variance in temperature, wind and rainfall in driving these trends (data that I collected for all sites and collections). Studies of mosquito-associated viruses have also found habitat disturbance to influence host community composition, and viral prevalence (Hermanns *et al.* 2021). It would also be interesting to investigate the influence of such anthropogenic factors in driving *Drosophila* virus prevalence, such as human population density, and green space. By characterising the environmental, and anthropogenic drivers of insect virus prevalence in temperate systems, we could also gain insight into how future climate change will impact insect virus prevalence. This will help to anticipate, and mitigate, the impacts of climate change on insects with economic and public health importance (Baylis 2017; Folly *et al.* 2020).

5.3 The evolution of Drosophila RNA viruses in multi-host, multi-virus systems

In this thesis, I was able to make some initial observations about the distribution of viral genotypes across *Drosophila* host species. In particular, for some closely related, multi-host viruses, such as Prestney burn, and Grom virus (Webster *et al.* 2016), I used sanger sequencing of regions of these viral genomes to confirm that these viruses were associated with multiple host species, with no obvious host-specific genotypic differentiation (Ch2). However, sequencing of more virus genes, and host individuals, would allow the quantification of the fidelity of *Drosophila* viruses to their host species, and the frequency of host switching, through Bayesian analyses of the heterochronous sequence data. To build a wider range of tip dates to temporally calibrate these phylogenies, we could use *Drosophila* samples collected in this system from Medd (2019, thesis). Quantifying the rates of host switching in a greater diversity of *Drosophila* viruses will allow us to model patterns of transmission between common and rare *Drosophila* species, and between generalist and specialists. I expect to find higher transmission from common host species to rare ones, between closely related species, and between those sharing resources, based on analyses of cross species transmission in vertebrate viruses (Faria *et al.* 2013; Luis *et al.* 2015). We might also be able to examine these sequences to determine whether any sites in the virus genomes are subject to recurrent positive selection – as previously seen in three codons in the *D.melanogaster* Nora virus capsid protein (Webster *et al.* 2015).

Finally, I see one of the possibly more speculative, but most exciting, possibilities for the spatiotemporally structured dataset collected in chapter 2 as the sequencing of single flies. As evidenced by Batson *et al.* (2020), single host sequencing can help us to better characterise rates of co-infection, prevalence, and viral diversity associated with insects. This study sequenced a total of 148 individual mosquitoes of seven species, collected from five habitats during autumn 2017. In contrast, the dataset I collected in chapter 2 represents a smaller spatial range, but monthly sampling over 3 years, and 15 *Drosophila* species. I kept 416 flies

as individuals for RNA extraction and virus assays, which represent every species-by-month by-site combination in the dataset. An exciting, but currently expensive, use of this dataset would be to conduct total RNA sequencing of each of these single flies, to obtain a heterochronous, and spatially structured dataset of drosophilids and their pathogens. As with the sanger sequencing, these samples could be combined with samples collected as part of Webster *et al.* (2016) and Medd (2019) to create a longer time series. With this data, not only could we characterise the total diversity of viruses present in Scottish *Drosophila* across years and seasons, but we could also characterise their rates of mutation, recombination, and reassortment, and historical patterns of host shifting, in a similar manner to studies of vertebrate viruses (eg. Worobey *et al.* 2020).

5.4 Drosophila DNA virus prevalence, and diversity

In comparison to RNA viruses of *Drosophila*, DNA viruses appear to be rare, and few have been identified (16 so far in Unkless 2011; Webster *et al.* 2015; Kapun *et al.* 2020; Wallace *et al.* 2020). However, they can recurrently evolve high virulence in the wild (Hill & Unckless 2020), and can suppress the host immune response (Palmer *et al.* 2018a). This suggests that they could be involved in co-evolutionary interactions with their hosts. However, due to the rarity of *Drosophila* DNA viruses, particularly in European *Drosophila*, the factors that predict their prevalence, and diversity were unknown. In chapter 3 of this thesis. I aimed to redress this imbalance by characterising the predictors of prevalence, and the patterns of genetic diversity seen in DNA viruses infecting European *Drosophila*. I found that the prevalence of three *Drosophila* DNA viruses (one double-stranded, one single-stranded and one single-stranded segmented) showed differing levels of spatial and temporal predictability. Population differentiation was high for all three viruses. However, I found that nucleotide diversity varies significantly, with the double-stranded DNA virus, Kallithea virus, showing much lower nucleotide diversity, and less constraint on sequence evolution in comparison to the single-stranded DNA viruses. This is consistent with its larger genome size, and the lower mutation

rate of double stranded DNA viruses (Williams *et al.* 2017). I found no evidence of positive selection on the Kallithea virus genome, in contrast with its closely relative (*Drosophila* innubila Nudivirus) which displays recurrent selection on host interacting proteins (Hill & Unckless 2018a). Whether this is due to the rarity of Kallithea virus, or some other inherent characteristic of its interaction with *Drosophila* hosts, would be an interesting subject for future studies.

One key aspect of the evolution of DNA viruses that I was unable to examine in this study is the frequency and effect of recombination. For eukaryotic ssDNA viruses, pervasive recombination could be a key driver of diversity, such as observed in canine parvovirus (Martin *et al.* 2011). This could potentially rescue genomic areas of low diversity induced by genetic bottlenecks, and explain some of the higher levels of diversity I saw in the ssDNA viruses. In the case of large dsDNA viruses, like the *Nudiviridae*, evidence from the closely related Baculoviruses suggests that they recombine frequently when co-infecting a single cell (Hajós *et al.* 2000; Cory & Myers 2003). This could allow recombination of adaptive mutations in the same genetic background, and facilitate more complex evolutionary dynamics. Future studies could investigate the frequency of DNA virus recombination, and re-assortment for segmented viruses, in individually sequenced wild flies, where single haplotypes would be easier to characterise.

5.5 The fitness costs of naturally-occurring Drosophila viruses

In chapter four, I characterised the fitness costs of infection with naturally-occurring *Drosophila* viruses, without the need for a viral isolate. I found that some *Drosophila* viruses readily transmit across species on (high-density) contact alone. This agrees with the wide ranging host generalism seen in my wild collected data (Ch2). I identified four viruses which significantly reduce Dmel lifespan, and three that significantly reduce Dmel total offspring production. Dmel Nora virus appears to not only reduce total offspring production, but the rate of offspring production in Dmel. Future studies could attempt to isolate the mechanism of this reduction in offspring production, and lifespan, and search for polymorphisms associated with

resistance in the hosts. The other three viruses with significant fitness costs to the host (Prestney burn virus, DimmNV and Muthill virus) do not have lab isolates (first described in van Mierlo *et al.* 2014; Webster *et al.* 2016). I characterise their possible effects for the first time, and demonstrate a method for this characterisation which does not depend on systemic injection. A more controlled version of this method, to increase sample size, could be used to elucidate the fitness effects of other *Drosophila* viruses without lab isolates (eg. Mondotte *et al.* 2018).

Additionally, I found not only that co-infection is common in the wild (Ch2), but also, in this experiment I found that the number of viral infections can decrease fly lifespan in a linear manner. This suggests that co-infections play a significant role in the ecological, and evolutionary dynamics of insect-virus interactions, and that further studies should examine whether infection with multiple viruses facilitates (eg. Kuwata *et al.* 2015) or restricts (eg. Vazeille *et al.* 2016) *Drosophila* virus replication and virulence. The sequential nature of wild co-infections may play a significant role in determining their outcome (Marchetto & Power 2017), as described in plant viruses. Alternatively, other factors such as the genetic similarity of co-infecting strains may influence their ability to establish in a host, as recently described in deformed wing virus (Gusachenko *et al.* 2021).

5.6 Conclusions

In this thesis, I have used drosophilids and their naturally occurring viruses to further contextualise the evolutionary genetics of invertebrate-virus interactions. I found that insect virus prevalence can vary significantly across host species, and season, providing evidence that ecological dynamics can drive inconsistency in insect virus-driven selection pressure. I also found that most viruses in this complex system infect multiple host species, enabling further studies of the determinants of this generalism, and the relationship between host species relatedness, and virus presence/absence. In the future, we should continue to work on understanding insect virus co-evolutionary interactions by studying them in the multi-host, multi-virus systems that they occur.

6 References

Aguiar, E.R.G.R., de Almeida, J.P.P., Queiroz, L.R., Oliveira, L.S., Olmo, R.P., da Silva De Faria, I.J., *et al.* (2020). A single unidirectional piRNA cluster similar to the flamenco locus is the major source of EVE-derived transcription and small RNAs in Aedes aegypti mosquitoes. *RNA*, 26, 581–594.

Albery, G.F., Eskew, E.A., Ross, N. & Olival, K.J. (2020). Predicting the global mammalian viral sharing network using phylogeography. *Nat. Commun.*, 11.

Altschul, S.F. (1990). Basic local alignment search tool. *J. Mol. Biol.*, 215.3, 403-410.

Ambrose, R.L., Lander, G.C., Maaty, W.S., Bothner, B., Johnson, J.E. & Johnson, K.N. (2009). Drosophila A virus is an unusual RNA virus with a T=3 icosahedral core and permuted RNA-dependent RNA polymerase. *J. Gen. Virol.*, 90, 2191–2200.

Andersen, K.G., Rambaut, A., Lipkin, W.I., Holmes, E.C. & Garry, R.F. (2020). The proximal origin of SARS-CoV-2. *Nat. Med.*, 26, 450–452.

Anderson, N., Adams, R.H., Demuth, J.P. & Blackmon, H. (2019). evobiR: Evolutionary Biology in R. R package version 1.3.

Anderson, R.M. & May, R.M. (1980). Infectious diseases and population cycles of forest insects. *Science*, 210, 658–661.

Arnold, P.A., Johnson, K.N. & White, C.R. (2013a). Physiological and metabolic consequences of viral infection in Drosophila melanogaster. *J. Exp. Biol.*, 216, 3350–7.

Arnold, P.A., Johnson, K.N. & White, C.R. (2013b). Physiological and metabolic consequences of viral infection in Drosophila melanogaster. *J. Exp. Biol.*, 216, 3350–3357.

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., *et al.* (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Curr. Protoc. Bioinforma.*, 43, 11.10.1-11.10.33.

Babayan, S.A., Orton, R.J. & Streicker, D.G. (2018). Predicting reservoir hosts and arthropod vectors from evolutionary signatures in RNA virus genomes. *Science (80-.).*, 362, 577–580.

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

Ballinger, M.J., Bruenn, J.A., Hay, J., Czechowski, D. & Taylor, D.J. (2014). Discovery and Evolution of Bunyavirids in Arctic Phantom Midges and Ancient Bunyavirid-Like Sequences in Insect Genomes. *J. Virol.*, 88, 8783 LP – 8794.

Ballinger, M.J., Bruenn, J.A. & Taylor, D.J. (2012). Phylogeny, integration and expression of sigma virus-like genes in Drosophila. *Mol. Phylogenet. Evol.*, 65, 251–258.

Basden, E.B. (1954). The distribution and biology of Drosophilidae (Diptera) in Scotland, including a new species of Drosophila . *Trans.Roy.Soc.Edin.*, 62, 603–654.

Batson, J., Dudas, G., Haas-Stapleton, E., Kistler, A.L., Li, L.M., Logan, P., *et al.* (2020).

Single mosquito metatranscriptomics recovers mosquito species, blood meal sources, and microbial cargo, including viral dark matter. *bioRxiv*, 2020.02.10.942854.

Baylis, M. (2017). Potential impact of climate change on emerging vector-borne and other infections in the UK. *Environ. Heal.*, 16, 112.

Behrman, E.L., Watson, S.S., O'Brien, K.R., Heschel, M.S. & Schmidt, P.S. (2015). Seasonal variation in life history traits in two Drosophila species. *J. Evol. Biol.*, 28, 1691–1704.

Bekal, S., Domier, L.L., Niblack, T.L. & Lambert, K.N. (2011). Discovery and initial analysis of novel viral genomes in the soybean cyst nematode. *J. Gen. Virol.*, 92, 1870–1879.

Benes, V., Blake, J. & Doyle, K. (2011). Ribo-Zero Gold Kit: improved RNA-seq results after removal of cytoplasmic and mitochondrial ribosomal RNA. *Nat. Methods*, 8.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B*, 57, 289–300.

Beraldo, C.S. (2018). Variação na susceptibilidade do hospedeiro a diferentes patógenos : um estudo experimental e filogenético de Drosophila -vírus.

Berkaloff, A., Bregliano, J.C. & Ohanessian, A. (1965). Mise en evidence de virions dans des drosophiles infectees par le virus hereditaire sigma. *Comptes Rendus Hebd. Des Seances L Acad. Des Sci.*, 5956–9.

Bigot, D., Atyame, C.M., Weill, M., Justy, F., Herniou, E.A. & Gayral, P. (2018). Discovery of Culex pipiens associated tunisia virus: a new ssRNA(+) virus representing a new insect associated virus family. *Virus Evol.*, 4, vex040.

Blangiardo, M., Cameletti, M., Baio, G. & Rue, H. (2013). Spatial and spatio-temporal models with R-INLA. *Spat. Spatiotemporal. Epidemiol.*

BLAST®, C.L.A.U.M. [Internet]. B. (MD): N.C. for B.I. (2008). Building a BLAST database with local sequences. In: *https://www.ncbi.nlm.nih.gov/books/NBK279688/*. Bethesda (MD).

Bolger, A.M., Lohse, M. & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30, 2114–2120.

Bombin, A. & Reed, L.K. (2016). The changing biodiversity of Alabama *Drosophila* : important impacts of seasonal variation, urbanization, and invasive species. *Ecol. Evol.*, 6, 7057–7069.

Bronkhorst, A.W., Van Cleef, K.W.R., Venselaar, H. & Van Rij, R.P. (2014). A dsRNA-binding protein of a complex invertebrate DNA virus suppresses the Drosophila RNAi response. *Nucleic Acids Res.*, 42, 12237–12248.

Brun, G., Plus, N., Ashburner, M. & Wright, T. (1980). *The viruses of Drosophila. In "The genetics and biology of Drosophila." New York Acad. Press*.

Buchfink, B., Xie, C. & Huson, D.H. (2014). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, 12, 59–60.

Buchon, N., Silverman, N. & Cherry, S. (2014). Immunity in Drosophila melanogaster - from microbial recognition to whole-organism physiology. *Nat. Rev. Immunol.*, 14, 796–810.

Calisher, C.H. & Gould, E.A. (2003). Taxonomy of the virus family Flaviviridae. *Adv. Virus*

Res., 59, 1–19.

Cao, C., Magwire, M.M., Bayer, F. & Jiggins, F.M. (2016). A Polymorphism in the Processing Body Component Ge-1 Controls Resistance to a Naturally Occurring Rhabdovirus in Drosophila. *PLoS Pathog.*, 12, 1–21.

Carlson, J., Suchman, E. & Buchatsky, L.B.T.-A. in V.R. (2006). Densoviruses for Control and Genetic Manipulation of Mosquitoes. In: *Insect Viruses: Biotechnological Applications*. Academic Press, pp. 361–392.

Caron, A., Bourgarel, M., Cappelle, J., Liégeois, F., De Nys, H.M. & Roger, F. (2018). Ebola virus maintenance: if not (Only) bats, what else? *Viruses*, 10.

Carrau, T., Hiebert, N., Vilcinskas, A. & Lee, K.-Z. (2018). Identification and characterisation of natural viruses associated with the invasive insect pest Drosophila suzukii. *J. Invertebr. Pathol.*

Chappell, T.M. & Rausher, M.D. (2016). Evolution of host range in Coleosporium ipomoeae, a plant pathogen with multiple hosts. *Proc. Natl. Acad. Sci. U. S. A.*, 113, 5346–5351.

Charif, D. & Lobry, J.R. (2007). SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis BT - Structural Approaches to Sequence Evolution: Molecules, Networks, Populations. In: (eds. Bastolla, U., Porto, M., Roman, H.E. & Vendruscolo, M.). Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 207–232.

Chiapello, M., Rodríguez-Romero, J., Ayllón, M.A. & Turina, M. (2020). Analysis of the virome associated to grapevine downy mildew lesions reveals new mycovirus lineages. *Virus Evol.*, 6.

Chouin-Carneiro, T., Vega-Rua, A., Vazeille, M., Yebakima, A., Girod, R., Goindin, D., *et al.* (2016). Differential susceptibilities of Aedes Aegypti and Aedes Albopictus from the Americas to Zika virus. *PLoS Negl Trop Dis*, 10.

Christian, P.D. (1987). Studies of Drosophila C and A viruses in Australian populations of Drosophila melanogaster, 305.

Chtarbanova, S., Lamiable, O., Lee, K.-Z., Galiana, D., Troxler, L., Meignin, C., *et al.* (2014). Drosophila C virus systemic infection leads to intestinal obstruction. *J. Virol.*

Ciota, A.T., Styer, L.M., Meola, M.A. & Kramer, L.D. (2011). The costs of infection and resistance as determinants of West Nile virus susceptibility in Culex mosquitoes. *BMC Ecol.*, 11, 23.

Cogni, R., Kuczynski, C., Koury, S., Lavington, E., Behrman, E.L., O'Brien, K.R., *et al.* (2014). THE INTENSITY OF SELECTION ACTING ON THE COUCH POTATO GENE—SPATIAL–TEMPORAL VARIATION IN A DIAPAUSE CLINE. *Evolution (N. Y).*, 68, 538–548.

Colmant, A.M.G., Etebari, K., Webb, C.E., Ritchie, S.A., Jansen, C.C., van den Hurk, A.F., *et al.* (2017a). Discovery of new orbiviruses and totivirus from Anopheles mosquitoes in Eastern Australia. *Arch. Virol.*, 162, 3529–3534.

Colmant, A.M.G., Hobson-Peters, J., Bielefeldt-Ohmann, H., van den Hurk, A.F., Hall-Mendelin, S., Chow, W.K., *et al.* (2017b). A New Clade of Insect-Specific Flaviviruses from Australian Anopheles Mosquitoes Displays Species-Specific Host Restriction. *mSphere*, 2, e00262-17.

Cordes, E.J., Licking-Murray, K.D. & Carlson, K.A. (2013). Differential gene expression related to Nora virus infection of Drosophila melanogaster. *Virus Res.*, 175, 95–100.

Cory, J.S. & Myers, J.H. (2003). The Ecology and Evolution of Insect Baculoviruses. *Annu. Rev. Ecol. Evol. Syst.*, 34, 239–272.

Cross, S.T., Maertens, B.L., Dunham, T.J., Rodgers, C.P., Brehm, A.L., Miller, M.R., *et al.* (2020). Partitiviruses Infecting Drosophila melanogaster and Aedes aegypti Exhibit Efficient Biparental Vertical Transmission . *J. Virol.*, 94, 1–17.

Dasgupta, R. & Sgro, J.-Y. (1989). *Nudeotide sequences of three Nodavinis RNA2's: the messengers for their coat protein precursors*. *Nucleic Acids Res.*

Davies, T.J. & Pedersen, a. B. (2008). Phylogeny and geography predict pathogen community similarity in wild primates and humans. *Proc. R. Soc. B Biol. Sci.*, 275, 1695–1701.

Delwart, E. & Li, L. (2012). Rapidly expanding genetic diversity and host range of the Circoviridae viral family and other Rep encoding small circular ssDNA genomes. *Virus Res.*, 164, 114–121.

Dinan, A.M., Lukhovitskaya, N.I., Olendraite, I. & Firth, A.E. (2020). A case for a negative-strand coding sequence in a group of positive-sense RNA viruses. *Virus Evol.*, 6.

Drummond, A.J. & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.*, 7, 214.

Duffy, S. (2018). Why are RNA virus mutation rates so damn high? *PLOS Biol.*, 16, e3000003.

Duffy, S., Shackelton, L.A. & Holmes, E.C. (2008). Rates of evolutionary change in viruses: patterns and determinants. *Nat. Rev. Genet.*, 9, 267–276.

Ekström, J.O. & Hultmark, D. (2016). A Novel Strategy for Live Detection of Viral Infection in Drosophila melanogaster. *Sci. Rep.*, 6.

F VogelsID, C.B., Rü ckertID, C., CavanyID, S.M., Alex PerkinsID, T., Ebel, G.D. & GrubaughID, N.D. (2019). Arbovirus coinfection and co-transmission: A neglected public health concern?

Faria, N.R., Suchard, M. a, Rambaut, A., Streicker, D.G. & Lemey, P. (2013). Simultaneously reconstructing viral cross-species transmission history and identifying the underlying constraints. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, 368, 20120196.

Fei, T., Ng, F., Willner, D.L., Lim, Y.W., Schmieder, R., Chau, B., *et al.* (2011). Broad Surveys of DNA Viral Diversity Obtained through Viral Metagenomics of Mosquitoes, 6.

Fenton, A. & Pedersen, A.B. (2005). Community epidemiology framework for classifying disease threats. *Emerg. Infect. Dis.*, 11, 1815–1821.

Fenton, A., Streicker, D.G., Petchey, O.L. & Pedersen, A.B. (2015). Are All Hosts Created Equal? Partitioning Host Species Contributions to Parasite Persistence in Multihost Communities. *Am. Nat.*, 186, 610–622.

Ferreira, Á.G., Naylor, H., Esteves, S.S., Pais, I.S., Martins, N.E. & Teixeira, L. (2014). The Toll-Dorsal Pathway Is Required for Resistance to Viral Oral Infection in Drosophila. *PLoS Pathog.*, 10.

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E. & Mello, C.C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, 391, 806–811.

Fleuriet, A. (1981). Comparison of various physiological traits in flies (Drosophila melanogaster) of wild origin, infected or uninfected by the hereditary Rhabdovirus sigma. *Arch. Virol.*, 69, 261–272.

Fleuriet, A. (1988). Maintenance of a Hereditary Virus. In: *Evolutionary Biology: Volume 23*. pp. 1–30.

Flt~uriet, A. (1982). *Archives of Virology Transmission Efficiency of the Siflnm Virus in Natural Populations of Its Host, Drosophila melanogaster*. *Axchives Virol.*

Folly, A.J., Dorey-Robinson, D., Hernández-Triana, L.M., Phipps, L.P. & Johnson, N. (2020). Emerging Threats to Animals in the United Kingdom by Arthropod-Borne Diseases. *Front. Vet. Sci.*, 7, 20.

Frangeul, L., Blanc, H., Saleh, M.C. & Suzuki, Y. (2020). Differential small RNA responses against co-infecting insect-specific viruses in aedes albopictus mosquitoes. *Viruses*, 12.

Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*.

Fujita, R., Kuwata, R., Kobayashi, D., Bertuso, A.G., Isawa, H. & Sawabe, K. (2017). Bustos virus, a new member of the negevirus group isolated from a Mansonia mosquito in the Philippines. *Arch. Virol.*, 162, 79–88.

Galbraith, D.A., Fuller, Z.L., Ray, A.M., Brockmann, A., Frazier, M., Gikungu, M.W., *et al.* (2018). Investigating the viral ecology of global bee communities with high-throughput metagenomics. *Sci. Rep.*, 8, 1–11.

Ge, L., Zhang, J., Zhou, X. & Li, H. (2007). Genetic Structure and Population Variability of Tomato Yellow Leaf Curl China Virus. *J. Virol.*, 81, 5902 LP – 5907.

Geden, C.J., Lietze, V.-U. & Boucias, D.G. (2008). *POPULATION BIOLOGY/GENETICS Seasonal Prevalence and Transmission of Salivary Gland Hypertrophy Virus of House Flies (Diptera: Muscidae)*. *J. Med. Entomol*.

Geoghegan, J.L., Duchêne, S. & Holmes, E.C. (2017). Comparative analysis estimates the relative frequencies of co-divergence and cross-species transmission within viral families. *PLoS Pathog.*, 13.

Geoghegan, J.L., Walker, P.J., Duchemin, J.-B. & Holmes, J.I. (2014). Seasonal Drivers of the Epidemiology of Arthropod-Borne Viruses in Australia. *PLoS Negl Trop Dis*, 8, 3325.

Geospiza. (2004). FinchTV.

Ginestet, C. (2011). ggplot2: Elegant Graphics for Data Analysis. *J. R. Stat. Soc. Ser. A (Statistics Soc.*, 174, 245–246.

Göertz, G.P., F Vogels, C.B., Geertsema, C., M Koenraadt, C.J. & Pijlman, G.P. (2017). Mosquito co-infection with Zika and chikungunya virus allows simultaneous transmission without affecting vector competence of Aedes aegypti.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., *et al.* (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome.

Nat. Biotechnol., 29, 644–652.

Greninger, A.L. (2018). A decade of RNA virus metagenomics is (not) enough. *Virus Res.*

Griffith, D.M., Veech, J.A. & Marsh, C.J. (2016). cooccur: Probabilistic Species Co-Occurrence Analysis in R. *J. Stat. Software; Vol 1, Code Snippet 2* .

Grozinger, C.M. & Flenniken, M.L. (2019). Bee Viruses: Ecology, Pathogenicity, and Impacts. *Annu. Rev. Entomol.*, 205–226.

Gupta, V., Stewart, C.O., Rund, S.S.C., Monteith, K. & Vale, P.F. (2017a). Costs and benefits of sublethal Drosophila C virus infection. *J. Evol. Biol.*, 30, 1325–1335.

Gupta, V., Stewart, C.O., Rund, S.S.C., Monteith, K. & Vale, P.F. (2017b). Costs and benefits of sublethal Drosophila C virus infection. *J. Evol. Biol.*, 30, 1325–1335.

Gusachenko, O.N., Woodford, L., Balbirnie-Cumming, K. & Evans, D.J. (2021). *First come, first served: Superinfection exclusion in Deformed wing virus is dependent upon sequence identity and not the order of virus acquisition*. *bioRxiv*. Available at: https://doi.org/10.1101/2021.03.22.436467. Last accessed .

Habayeb, M.S., Cantera, R., Casanova, G., Ekström, J.-O., Albright, S. & Hultmark, D. (2009a). The Drosophila Nora virus is an enteric virus, transmitted via feces. *J. Invertebr. Pathol.*, 101, 29–33.

Habayeb, M.S., Ekengren, S.K. & Hultmark, D. (2006). Nora virus, a persistent vitus in Drosophila, defines a new picorna-like virus family. *J. Gen. Virol.*, 87, 3045–3051.

Habayeb, M.S., Ekström, J.-O. & Hultmark, D. (2009b). Nora Virus Persistent Infections Are Not Affected by the RNAi Machinery. *PLoS One*, 4, e5731.

Hadfield, J.D. (2010). MCMCglmm: MCMC Methods for Multi-Response GLMMs in R. *J. Stat. Softw.*, 33, 1–22.

Hadfield, J.D., Krasnov, B.R., Poulin, R. & Nakagawa, S. (2014). A tale of two phylogenies: Comparative analyses of ecological interactions. *Am. Nat.*

Hajós, J.P., Pijnenburg, J., Usmany, M., Zuidema, D., Závodszky, P. & Vlak, J.M. (2000). High frequency recombination between homologous baculoviruses in cell culture. *Arch. Virol.*, 145, 159–164.

Hall, A.R., Ashby, B., Bascompte, J. & King, K.C. (2020). Measuring Coevolutionary Dynamics in Species-Rich Communities. *Trends Ecol. Evol.*

Hall, T.. (1999). BioEdit: A User-Friendly Biological Sequence Alignment Editor and Analysis Program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.*, 95–98.

Hannon, G.J. (2010). FASTX-Toolkit.

Hedges, L.M. & Johnson, K.N. (2008). Induction of host defence responses by Drosophila C virus. *J. Gen. Virol.*

Hermanns, K., Marklewitz, M., Zirkel, F., Kopp, A., Kramer-Schadt, S. & Junglen, S. (2021). Mosquito community composition shapes virus prevalence patterns along anthropogenic disturbance gradients. *bioRxiv*.

Hill, T. & Unckless, R.L. (2017). Baculovirus Molecular Evolution via Gene Turnover and Recurrent Positive Selection of Key Genes. *J. Virol.*, 91, e01319-17.

Hill, T. & Unckless, R.L. (2018a). The dynamic evolution of Drosophila innubila Nudivirus. *Infect. Genet. Evol.*, 57, 151–157.

Hill, T. & Unckless, R.L. (2018b). The dynamic evolution of Drosophila innubila Nudivirus. *Infect. Genet. Evol.*, 57, 151–157.

Hill, T. & Unckless, R.L. (2020). Recurrent evolution of high virulence in isolated populations of a DNA virus. *Elife*, 9, e58931.

Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q. & Vinh, L.S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation. Molecular biology and evolution. *Mol. Biol. Evol.*, 35, 518–522.

Holt, R.A., Subramanian, G.M., Halpern, A., Sutton, G.G., Charlab, R., Nusskern, D.R., *et al.* (2002). The Genome Sequence of the Malaria Mosquito Anopheles gambiae *Science (80-.).*, 298, 129 LP – 149.

ter Horst, A.M., Nigg, J.C., Dekker, F.M. & Falk, B.W. (2019). Endogenous Viral Elements Are Widespread in Arthropod Genomes and Commonly Give Rise to PIWI-Interacting RNAs. *J. Virol.*, 93, e02124-18.

Imrie, R., Roberts, K.E. & Longdon, B. (2021). Between virus correlations in the outcome of infection across host species: evidence of virus genotype by host species interactions. *bioRxiv*, 2021.02.16.431403.

Institute, B. (2020). Picard Tools.

Joosten, J., van Rij, R.P. & Miesen, P. (2020). Slicing of viral RNA guided by endogenous piRNAs triggers the production of responder and trailer piRNAs in Aedes mosquitoes. *bioRxiv*.

Jousset, F.-X., Bergoin, M. & Revet, B. (1977). Characterization of the Drosophila C virus. *J. Gen. Virol.*

Jousset, F., Plus, N., Croizier, G. & Thomas, M. (1972). Existence in Drosophila of 2 groups of picornavirus with different biological and serological properties. *C R Acad Sci Hebd Seances Acad Sci D.*, 725, 3043–3046.

Junglen, S., Korries, M., Grasse, W., Wieseler, J., Kopp, A., Hermanns, K., *et al.* (2017). Host Range Restriction of Insect-Specific Flaviviruses Occurs at Several Levels of the Viral Life Cycle. *mSphere*, 2, e00375-16.

Käfer, S., Paraskevopoulou, S., Zirkel, F., Wieseke, N., Donath, A., Petersen, M., *et al.* (2019). Re-assessing the diversity of negative strand RNA viruses in insects. *PLOS Pathog.*, 15, e1008224.

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., Von Haeseler, A. & Jermiin, L.S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods*, 14, 587–589.

Kapun, M., Barron, M.G., Staubach, F., Obbard, D.J., Axel, R., Vieira, J., *et al.* (2020). Genomic analysis of european drosophila melanogaster populations reveals longitudinal structure, continent-wide selection, and previously unknown DNA viruses. *Mol. Biol. Evol.*, 37, 2661–2678.

Kapun, M., Nolte, V., Flatt, T., Schlö, C. & Welch, J.J. (2010). Host Range and Specificity of the Drosophila C Virus.

Kapun, M., Nunez, J.C.B., Bogaerts-márquez, M. & Murga-, J. (2021). Drosophila Evolution over Space and Time (DEST) - A New Population Genomics Resource. *biorxiv*.

Kariithi, H.M., Meki, I.K., Boucias, D.G. & Abd-Alla, A.M.M. (2017). Hytrosaviruses: current status and perspective. *Curr. Opin. Insect Sci.*, 22, 71–78.

Keele, B.F., Van Heuverswyn, F., Li, Y., Bailes, E., Takehisa, J., Santiago, M.L., *et al.* (2006). Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science*, 313, 523–6.

Kemp, C., Mueller, S., Goto, A., Barbier, V., Paro, S., Bonnay, F., *et al.* (2013). Broad RNA interference-mediated antiviral immunity and virus-specific inducible responses in Drosophila. *J Immunol*.

Kim, B.Y., Wang, J.R., Miller, D.E., Barmina, O., Delaney, E., Thompson, A., *et al.* (2020). Highly contiguous assemblies of 101 drosophilid genomes. *bioRxiv*, 2020.12.14.422775.

Kim, K.-H. & Bae, J.-W. (2011). Amplification Methods Bias Metagenomic Libraries of Uncultured Single-Stranded and Double-Stranded DNA Viruses. *Appl. Environ. Microbiol.*, 77, 7663 LP – 7668.

Kimenyi, K.M., Abry, M.F., Okeyo, W., Matovu, E., Masiga, D. & Kulohoma, B.W. (2020). Detecting bracoviral orthologs distribution in five tsetse fly species and the housefly genomes. *BMC Res. Notes*, 13, 318.

Kofler, R., Orozco-terWengel, P., de Maio, N., Pandey, R.V., Nolte, V., Futschik, A., *et al.* (2011a). Popoolation: A toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS One*, 6, e15925.

Kofler, R., Pandey, R.V. & Schlötterer, C. (2011b). PoPoolation2: Identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics*, 27, 3435–3436.

Kolaczkowski, B., Hupalo, D.N. & Kern, A.D. (2011). Recurrent Adaptation in RNA Interference Genes Across the Drosophila Phylogeny Research article, 28, 1033–1042.

Kolde, R. (2015). pheatmap: Pretty heatmaps [Software].

Krueger, F. (2015). Trim galore. *A wrapper tool around Cutadapt FastQC to consistently apply Qual. Adapt. trimming to FastQ files*, 516.

Kuwata, R., Isawa, H., Hoshino, K., Sasaki, T., Kobayashi, M., Maeda, K., *et al.* (2015). Analysis of Mosquito-Borne Flavivirus Superinfection in Culex tritaeniorhynchus (Diptera: Culicidae) Cells Persistently Infected with Culex Flavivirus (Flaviviridae). *J. Med. Entomol.*, 52, 222–229.

L'Heritier, P. & Teissier, G. (1937). Une anomalie physiologique herediaire chez la Drosophile. *CR Acad*.

Lamiable, O., Kellenberger, C., Kemp, C., Troxler, L., Pelte, N., Boutros, M., *et al.* (2016). Cytokine Diedel and a viral homologue suppress the IMD pathway in Drosophila. *Proc. Natl. Acad. Sci.*, 113, 698–703.

Langmead, B. & Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, 9, 357–359.

LeBeau, B. (2020). simglm: Simulate Models Based on the Generalized Linear Model. R

package version 0.8.0.

Levitt, A.L., Singh, R., Cox-Foster, D.L., Rajotte, E., Hoover, K., Ostiguy, N., *et al.* (2013). Cross-species transmission of honey bee viruses in associated arthropods. *Virus Res.*, 176, 232–240.

Li, C.-X., Shi, M., Tian, J.-H., Lin, X.-D., Kang, Y.-J., Chen, L.-J., *et al.* (2015a). Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *Elife*, 4, 1–26.

Li, C.X., Shi, M., Tian, J.H., Lin, X.D., Kang, Y.J., Chen, L.J., *et al.* (2015b). Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *Elife*, 4, 4–6.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*, 2.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., *et al.* (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25, 2078–2079.

Lietze, V.-U., Abd-Alla, A.M.M., Vreysen, M.J.B., Geden, C.J. & Boucias, D.G. (2011). Salivary gland hypertrophy viruses: a novel group of insect pathogenic viruses. *Annu. Rev. Entomol.*, 56, 63–80.

Lindgren, F. & Rue, H. (2015). Bayesian spatial modelling with R-INLA. *J. Stat. Softw.*, 63, 1–25.

Lobo, F.P., Mota, B.E.F., Pena, S.D.J., Azevedo, V., Macedo, A.M., Tauch, A., *et al.* (2009). Virus-host coevolution: Common patterns of nucleotide motif usage in Flaviviridae and their hosts. *PLoS One*, 4, 6282.

Long, J.S., Mistry, B., Haslam, S.M. & Barclay, W.S. (2019). Host and viral determinants of influenza A virus species specificity. *Nat. Rev. Microbiol.*, 17, 67–81.

Longdon, B., Brockhurst, M. a., Russell, C. a., Welch, J.J. & Jiggins, F.M. (2014). The Evolution and Genetics of Virus Host Shifts. *PLoS Pathog.*, 10, e1004395.

Longdon, B., Day, J.P., Alves, J.M., Smith, S.C.L., Houslay, T.M., Mcgonigle, J.E., *et al.* (2018). Host shifts result in parallel genetic changes when viruses evolve in closely related species.

Longdon, B., Day, J.P., Schulz, N., Leftwich, P.T., de Jong, M.A., Breuker, C.J., *et al.* (2017). Vertically transmitted rhabdoviruses are found across three insect families and have dynamic interactions with their hosts. *Proc. R. Soc. B Biol. Sci.*, 284, 20162381.

Longdon, B., Hadfield, J.D., Day, J.P., Smith, S.C.L., McGonigle, J.E., Cogni, R., *et al.* (2015a). The causes and consequences of changes in virulence following pathogen host shifts. *PLoS Pathog.*, 11, e1004728.

Longdon, B., Hadfield, J.D., Webster, C.L., Obbard, D.J. & Jiggins, F.M. (2011a). Host Phylogeny Determines Viral Persistence and Replication in Novel Hosts. *PLoS Pathog.*, 7, e1002260.

Longdon, B., Murray, G.G.R., Palmer, W.J., Day, J.P., Parker, D.J., Welch, J.J., *et al.* (2015b). The evolution, diversity, and host associations of rhabdoviruses. *Virus Evol.*, 1, 1–12.

Longdon, B., Obbard, D.J. & Jiggins, F.M. (2010). Sigma viruses from three species of

Drosophila form a major new clade in the rhabdovirus phylogeny. *Proc. R. Soc. B Biol. Sci.*, 277, 35–44.

Longdon, B., Wilfert, L., Obbard, D.J. & Jiggins, F.M. (2011b). Rhabdoviruses in two species of drosophila: Vertical transmission and a recent sweep. *Genetics*.

Longdon, B., Wilfert, L., Osei-Poku, J., Cagney, H., Obbard, D.J. & Jiggins, F.M. (2011c). Host-switching by a vertically transmitted rhabdovirus in Drosophila. *Biol. Lett.*, 7, 747–750.

Lopez, W., Page, A.M., Carlson, D.J., Ericson, B.L., Cserhati, M.F., Guda, C., *et al.* (2018). Analysis of immune-related genes during Nora virus infection of Drosophila melanogaster using next generation sequencing, 4, 123–139.

Löytynoja, A. (2014). Phylogeny-aware alignment with PRANK. *Methods Mol. Biol.*, 1079, 155–170.

Luis, A.D., O'Shea, T.J., Hayman, D.T.S., Wood, J.L.N., Cunningham, A. a., Gilbert, A.T., *et al.* (2015). Network analysis of host-virus communities in bats and rodents reveals determinants of cross-species transmission. *Ecol. Lett.*, n/a-n/a.

Lumme, J. & Laakovaara, S. (1983). Seaonality and Diapause in the Drosophilids. In: *The Genetics and Biology of Drosophila* (eds. Ashburner, M., Carson, H.L. & Thompson, J.N.). Academic Press, London.

Machado, H., Bergland, A., Taylor, R., Tilk, S., Behrman, E., Dyer, K., *et al.* (2019). Broad geographic sampling reveals predictable, pervasive, and strong seasonal adaptation in Drosophila. *bioRxiv*.

Magwire, M.M., Fabian, D.K., Schweyen, H., Cao, C., Longdon, B., Bayer, F., *et al.* (2012). Genome-Wide Association Studies Reveal a Simple Genetic Basis of Resistance to Naturally Coevolving Viruses in Drosophila melanogaster. *PLoS Genet.*, 8.

Mahar, J.E., Shi, M., Hall, R.N., Strive, T. & Holmes, E.C. (2020). Comparative Analysis of RNA Virome Composition in Rabbits and Associated Ectoparasites. *J. Virol.*, 94, e02119-19.

Manley, R., Boots, M. & Wilfert, L. (2015). REVIEW: Emerging viral disease risk to pollinating insects: ecological, evolutionary and anthropogenic factors. *J. Appl. Ecol.*, 52, 331–340.

Marchetto, K.M. & Power, A.G. (2017). Coinfection Timing Drives Host Population Dynamics through Changes in Virulence. *Am. Nat.*, 191, 173–183.

Martin, D.P., Biagini, P., Lefeuvre, P., Golden, M., Roumagnac, P. & Varsani, A. (2011). Recombination in Eukaryotic Single Stranded DNA Viruses. *Viruses*, 3, 1699–1738.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal; Vol 17, No 1 Next Gener. Seq. Data Anal. - 10.14806/ej.17.1.200* .

Martin, S.J., Highfield, A.C., Brettell, L., Villalobos, E.M., Budge, G.E., Powell, M., *et al.* (2012). Global Honey Bee Viral Landscape Altered by a Parasitic Mite. *J. Anim. Ecol.*, 336, 1304–1306.

Martins, N.E., Faria, V.G., Teixeira, L., Magalhães, S. & Sucena, É. (2013). Host Adaptation Is Contingent upon the Infection Route Taken by Pathogens. *PLoS Pathog.*, 9.

McMahon, D.P., Wilfert, L., Paxton, R.J. & Brown, M.J.F. (2018). Chapter Eight - Emerging Viruses in Bees: From Molecules to Ecology. In: *Environmental Virology and Virus Ecology* (ed. Malmstrom, C.M.B.T.-A. in V.R.). Academic Press, pp. 251–291.

Medd, N.C. (2019). Viruses and antiviral responses of an invasive fruit pest, Drosophila suzukii. University of Edinburgh.

Medd, N.C., Fellous, S., Waldron, F.M., Xue, A., Nakai, M., Cross, J. V, *et al.* (2018a). The virome of Drosophila suzukii, an invasive pest of soft fruit. *Virus Evol.*, 4, 1–14.

Medd, N.C., Fellous, S., Waldron, F.M., Xuéreb, A., Nakai, M., Cross, J. V., *et al.* (2018b). The virome of Drosophila suzukii, an invasive pest of soft fruit. *Virus Evol.*, 4.

Merkling, S.H. & van Rij, R.P. (2015). Analysis of resistance and tolerance to virus infection in Drosophila. *Nat. Protoc.*, 10.

Michalakis, Y. & Blanc, S. (2020). The Curious Strategy of Multipartite Viruses. *Annu. Rev. Virol.*, 7, 203–218.

van Mierlo, J.T., Overheul, G.J., Obadia, B., van Cleef, K.W.R., Webster, C.L., Saleh, M.C., *et al.* (2014). Novel Drosophila Viruses Encode Host-Specific Suppressors of RNAi. *PLoS Pathog.*, 10.

de Miranda, J.R., Hedman, H., Onorati, P., Stephan, J., Karlberg, O., Bylund, H., *et al.* (2017). Characterization of a Novel RNA Virus Discovered in the Autumnal Moth Epirrita autumnata in Sweden. *Viruses*, 9, 214.

Mlih, M., Khericha, M., Birdwell, C., West, A.P. & Karpac, J. (2018). A virus-acquired host cytokine controls systemic aging by antagonizing apoptosis. *PLOS Biol.*, 16, e2005796.

Mondet, F., de Miranda, J.R., Kretzschmar, A., Le Conte, Y. & Mercer, A.R. (2014). On the Front Line: Quantitative Virus Dynamics in Honeybee (Apis mellifera L.) Colonies along a New Expansion Front of the Parasite Varroa destructor. *PLoS Pathog.*, 10.

Mondotte, J.A., Gausson, V., Frangeul, L., Blanc, H., Lambrechts, L. & Saleh, M.C. (2018). Immune priming and clearance of orally acquired RNA viruses in Drosophila. *Nat. Microbiol.*, 3, 1394–1403.

Mondotte, J.A. & Saleh, M.C. (2018). Antiviral Immune Response and the Route of Infection in Drosophila melanogaster. *Adv. Virus Res.*, 100, 247–278.

Moreau, Y., Gil, P., Exbrayat, A., Rakotoarivony, I., Bréard, E., Sailleau, C., *et al.* (2020). The genome segments of Bluetongue virus differ in copy number in a host-specific manner. *J. Virol.*, JVI.01834-20.

Moriyasu, Y., Maruyama-Funatsuki, W., Kikuchi, A., Ichimi, K., Zhong, B., Yan, J., *et al.* (2007). Molecular analysis of the genome segments S1, S4, S6, S7 and S12 of a Rice gall dwarf virus isolate from Thailand; completion of the genomic sequence. *Arch. Virol.*, 152, 1315–1322.

Mussabekova, A., Daeffler, L. & Imler, J.L. (2017). Innate and intrinsic antiviral immunity in Drosophila. *Cell. Mol. Life Sci.*, 2039–2054.

Myer, M.H. & Johnston, J.M. (2019). Spatiotemporal Bayesian modeling of West Nile virus: Identifying risk of infection in mosquitoes with local-scale predictors. *Sci. Total Environ.*, 650, 2818–2829.

Natsopoulou, M.E., Mcmahon, D.P., Doublet, V., Bryden, J. & Paxton, R.J. (2015).

Interspecific competition in honeybee intracellular gut parasites is asymmetric and favours the spread of an emerging infectious disease. *Proc. R. Soc. B*, 282.

Nayak, A., Berry, B., Tassetto, M., Kunitomi, M., Acevedo, A., Deng, C., *et al.* (2010). Cricket paralysis virus antagonizes Argonaute 2 to modulate antiviral defense in Drosophila. *Nat. Struct. Mol. Biol.*, 17, 547–554.

Neyen, C., Bretscher, A.J., Binggeli, O. & Lemaitre, B. (2014). Methods to study Drosophila immunity. *Methods*, 68, 116–128.

Nguyen, L.T., Schmidt, H.A., Von Haeseler, A. & Minh, B.Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*, 32, 268–274.

Nunes, M.R.T., Contreras-Gutierrez, M.A., Guzman, H., Martins, L.C., Barbirato, M.F., Savit, C., *et al.* (2017). Genetic characterization, molecular epidemiology, and phylogenetic relationships of insect-specific viruses in the taxon Negevirus. *Virology*, 504, 152–167.

O'Brien, C.A., McLean, B.J., Colmant, A.M.G., Harrison, J.J., Hall-Mendelin, S., van den Hurk, A.F., *et al.* (2017). Discovery and Characterisation of Castlerea Virus, a New Species of Negevirus Isolated in Australia. *Evol. Bioinform. Online*, 13, 1176934317691269.

O'Brien, E. & Xagoraraki, I. (2019). Understanding temporal and spatial variations of viral disease in the US: The need for a one-health-based data collection and analysis approach. *One Heal. (Amsterdam, Netherlands)*, 8, 100105.

Obbard, D.J. & Dudas, G. (2014). The genetics of host-virus coevolution in invertebrates. *Curr. Opin. Virol.*

Obbard, D.J., Jiggins, F.M., Bradshaw, N.J. & Little, T.J. (2011). Recent and Recurrent Selective Sweeps of the Antiviral RNAi Gene Argonaute-2 in Three Species of Drosophila. *Mol. Biol. Evol.*, 28, 1043–1056.

Obbard, D.J., Jiggins, F.M., Halligan, D.L. & Little, T.J. (2006). Natural selection drives extremely rapid evolution in antiviral RNAi genes. *Curr. Biol.*, 16, 580–585.

Obbard, D.J., Shi, M., Roberts, K.E., Longdon, B. & Dennis, A.B. (2020). A new lineage of segmented RNA viruses infecting animals. *Virus Evol.*, 6.

Odindo, M.O. (1982). Incidence of Salivary Gland Hypertrophy in Field Populations of the Tsetse Glossina Pallidipes on the South Kenyan Coast. *Int. J. Trop. Insect Sci.*, 3, 59–64.

Odindo, M.O. & Amutalla, P.A. (1986). Distribution Pattern of the Virus of Glossina pallidipes Austen in a Forest Ecosystem. *Int. J. Trop. Insect Sci.*, 7, 79–84.

Oksanen, A.J., Blanchet, F.G., Kindt, R., Legen-, P., Minchin, P.R., Hara, R.B.O., *et al.* (2012). Community Ecology Package. *… Ecol. Packag. …*, 2, 263.

Olivier, V., Blanchard, P., Chaouch, S., Lallemand, P., Schurr, F., Celle, O., *et al.* (2008). Molecular characterisation and phylogenetic analysis of Chronic bee paralysis virus, a honey bee virus. *Virus Res.*, 132, 59–68.

Palacios, G., Savji, N., Travassos da Rosa, A., Guzman, H., Yu, X., Desai, A., *et al.* (2013). Characterization of the Uukuniemi Virus Group (Phlebovirus: Bunyaviridae): Evidence for Seven Distinct Species. *J. Virol.*, 87, 3187 LP – 3195.

Palatini, U., Miesen, P., Carballar-Lejarazu, R., Ometto, L., Rizzo, E., Tu, Z., *et al.* (2017a). Comparative genomics shows that viral integrations are abundant and express piRNAs in the arboviral vectors Aedes aegypti and Aedes albopictus. *BMC Genomics*, 18, 512.

Palatini, U., Miesen, P., Carballar-Lejarazu, R., Ometto, L., Rizzo, E., Tu, Z., *et al.* (2017b). Comparative genomics shows that viral integrations are abundant and express piRNAs in the arboviral vectors Aedes aegypti and Aedes albopictus. *BMC Genomics*, 18, 512.

Palmer, W.H., Joosten, J., Overheul, G.J., Jansen, P.W., Vermeulen, M., Obbard, D.J., *et al.* (2018a). Induction and Suppression of NF-κB Signalling by a DNA Virus of Drosophila. *J. Virol.*, 93, e01443-18.

Palmer, W.H., Medd, N.C., Beard, P.M. & Obbard, D.J. (2018b). Isolation of a natural DNA virus of Drosophila melanogaster, and characterisation of host resistance and immune responses. *PLoS Pathog.*, 14.

Parrish, C.R., Holmes, E.C., Morens, D.M., Park, E.-C., Burke, D.S., Calisher, C.H., *et al.* (2008). Cross-species virus transmission and the emergence of new epidemic diseases. *Microbiol. Mol. Biol. Rev.*, 72, 457–470.

Pauszek, S.J., Allende, R. & Rodriguez, L.L. (2008). Characterization of the full-length genomic sequences of vesicular stomatitis Cocal and Alagoas viruses. *Arch. Virol.*, 153, 1353–1357.

Pettersson, J.H.-O., Shi, M., Eden, J.-S., Holmes, E.C. & Hesson, J.C. (2019). Meta-Transcriptomic Comparison of the RNA Viromes of the Mosquito Vectors Culex pipiens and Culex torrentium in Northern Europe. *Viruses*.

Piégu, B., Guizard, S., Yeping, T., Cruaud, C., Asgari, S., Bideshi, D.K., *et al.* (2014). Genome sequence of a crustacean iridovirus, IIV31, isolated from the pill bug, Armadillidium vulgare. *J. Gen. Virol.*, 95, 1585–1590.

Plus, N., Croizier, G., Jousset, F.X. & David, J. (1975). Picornaviruses of laboratory and wild Drosophila melanogaster: geographical distribution and serotypic composition. *Ann. Microbiol. (Paris).*, 126, 107—117.

Plus, N. & Duthoit, J.. (1969). Un nouveau virus de Drosophila melanogaster, le virus P. *Comptes rendus l'Acaddmie des Sci.*, 268.

Porter, A.F., Shi, M., Eden, J.-S., Zhang, Y.-Z. & Holmes, E.C. (2019). Diversity and Evolution of Novel Invertebrate DNA Viruses Revealed by Meta-Transcriptomics. *Viruses*, 11, 1092.

Quinlan, A.R. & Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26, 841–842.

R Core Team. (2019). R: A language and environment for statistical computing. *R Found. Stat. Comput.*

R Core Team. (2020). R: A language and environment for statistical computing. *R Found. Stat. Comput. , Vienna, Austria.*

Rambaut, A., Drummond, A.J., Xie, D., Baele, G. & Suchard, M.A. (2018). Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst. Biol.*, 67, 901–904.

Ramsey, S.D., Ochoa, R., Bauchan, G., Gulbronson, C., Mowery, J.D., Cohen, A., *et al.* (2019). *Varroa destructor* feeds primarily on honey bee fat body tissue and not hemolymph. *Proc. Natl. Acad. Sci.*, 116, 1792–1801.

Ratcliff, F., Harrison, B.D. & Baulcombe, D.C. (1997). A Similarity Between Viral Defense and Gene Silencing in Plants. *Science (80-.).*, 276, 1558 LP – 1560.

Dos Reis, M., Hay, A.J. & Goldstein, R.A. (2009). Using non-homogeneous models of nucleotide substitution to identify host shift events: Application to the origin of the 1918 “spanish” influenza pandemic virus. *J. Mol. Evol.*, 69, 333–345.

Remnant, E.J. (2017). A Diverse Range of Novel RNA Viruses in Geographically Distinct Honey Bee Populations. *J. Virol.*, 91, 1–19.

Ribeiro, J.M., Debat, H.J., Boiani, M., Ures, X., Rocha, S. & Breijo, M. (2019). An insight into the sialome, mialome and virome of the horn fly, Haematobia irritans. *BMC Genomics*, 20, 616.

Rohwer, F. (2003). Global Phage Diversity. *Cell*, 113, 141.

Roossinck, M.J. (2011). The good viruses: viral mutualistic symbioses. *Nat. Rev. Microbiol.*, 9, 99–108.

Rosario, K., Mettel, K.A., Benner, B.E., Johnson, R., Scott, C., Yusseff-Vanegas, S.Z., *et al.* (2018). Virus discovery in all three major lineages of terrestrial arthropods highlights the diversity of single-stranded DNA viruses associated with invertebrates. *PeerJ*, 6, e5761.

Rosewell, J. & Shorrocks, B. (1987). The implication of survival rates in natural populations of Drosophila: capture-recapture experiments on domestic species. *Biol. J. Linn. Soc.*, 32, 373–384.

Rossignol, P.A., Ribeiro, J.M., Jungery, M., Turell, M.J., Spielman, A. & Bailey, C.L. (1985). Enhanced mosquito blood-finding success on parasitemic hosts: evidence for vector-parasite mutualism. *Proc. Natl. Acad. Sci. U. S. A.*, 82, 7725–7.

Ruark, C.L., Gardner, M., Mitchum, M.G., Davis, E.L. & Sit, T.L. (2018). Novel RNA viruses within plant parasitic cyst nematodes. *PLoS One*, 13, e0193881.

Runckel, C., Flenniken, M.L., Engel, J.C., Ruby, J.G., Ganem, D., Andino, R., *et al.* (2011). Temporal Analysis of the Honey Bee Microbiome Reveals Four Novel Viruses and Seasonal Prevalence of Known Viruses, Nosema, and Crithidia. *PLoS One*, 6.

Russo, C.A., Takezaki, N. & Nei, M. (1995). Molecular phylogeny and divergence times of drosophilid species. *Mol. Biol. Evol.*, 12, 391–404.

Schultz, M.J., Frydman, H.M. & Connor, J.H. (2018). Dual Insect specific virus infection limits Arbovirus replication in Aedes mosquito cells. *Virology*, 518, 406–413.

Sedger, L., Collins, M.H., Hughes, G.L., Robin, C., Maciel-De-Freitas, R., Dias Da Silveira, I., *et al.* (2018). Zika Virus Infection Produces a Reduction on Aedes aegypti Lifespan but No Effects on Mosquito Fecundity and Oviposition Success, 9, 3011.

Seecof, R.. (1964). Deletarious effects on Drosophila development associated with sigma virus infection. *Virology*, 142–148.

Shackelton, L.A., Parrish, C.R., Truyen, U. & Holmes, E.C. (2005). High rate of viral evolution associated with the emergence of carnivore parvovirus. *Proc. Natl. Acad. Sci. U. S. A.*, 102, 379–84.

Shapiro, B., Rambaut, A. & Drummond, A.J. (2006). Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Mol. Biol. Evol.*

Sharpe, S.R., Morrow, J.L., Brettell, L.E., Shearman, D.C., Gilchrist, S., Cook, J.M., *et al.* (2021). Tephritid fruit flies have a large diversity of co-occurring RNA viruses. *J. Invertebr. Pathol.*

Shi, M., Lin, X.-D., Tian, J.-H., Chen, L.-J., Chen, X., Li, C.-X., *et al.* (2016a). Redefining the invertebrate RNA virosphere. *Nature*.

Shi, M., Lin, X.-D., Tian, J.-H., Chen, L.-J., Chen, X., Li, C.-X., *et al.* (2016b). Redefining the invertebrate RNA virosphere. *Nature*, 1–12.

Shi, M., Lin, X.-D., Vasilakis, N., Tian, J.-H., Li, C.-X., Chen, L.-J., *et al.* (2016c). Divergent Viruses Discovered in Arthropods and Vertebrates Revise the Evolutionary History of the Flaviviridae and Related Viruses. *J. Virol.*, 90, 659–669.

Shi, M., Neville, P., Nicholson, J., Eden, J.-S., Imrie, A. & Holmes, E.C. (2017). High-Resolution Metatranscriptomics Reveals the Ecological Dynamics of Mosquito-Associated RNA Viruses in Western Australia. *J. Virol.*, 91, 1–17.

Shi, M., White, V.L., Schlub, T., Eden, J.-S., Hoffmann, A.A. & Holmes, E.C. (2018). No detectable effect of Wolbachia wMel on the prevalence and abundance of the RNA virome of Drosophila melanogaster. *Proc. R. Soc. B Biol. Sci.*, 285, 20181165.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., *et al.* (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, 7, 539.

Signor, S.A., New, F.N. & Nuzhdin, S. (2018). A Large Panel of Drosophila simulans Reveals an Abundance of Common Variants. *Genome Biol. Evol.*, 10, 189–206.

Speybroeck, N., Williams, C., Lafia, K., Devleesschauwer, B. & Berkvens, D. (2012). Estimating the prevalence of infections in vector populations using pools of samples. *Med. Vet. Entomol.*, 26.

Stenger, D.C., Sisterson, M.S. & French, R. (2010). Population genetics of Homalodisca vitripennis reovirus validates timing and limited introduction to California of its invasive insect host, the glassy-winged sharpshooter. *Virology*, 407, 53–59.

Streicker, D.G., Turmelle, a S., Vonhof, M.J., Kuzmin, I. V, McCracken, G.F. & Rupprecht, C.E. (2010). Host phylogeny constrains cross-species emergence and establishment of rabies virus in bats. *Science (80-.).*, 329, 676–679.

Styer, L.M., Meola, M.A. & Kramer, L.D. (2007). West Nile Virus Infection Decreases Fecundity of Culex tarsalis Females. *J. Med. Entomol.*, 44, 1074–1085.

Suvorov, A., Kim, B.Y., Wang, J., Armstrong, E.E., Peede, D., D'Agostino, E.R.R., *et al.* (2021). Widespread introgression across a phylogeny of 155 Drosophila genomes. *bioRxiv*, 2020.12.14.422758.

Takeda, K. & Akira, S. (2005). Toll-like receptors in innate immunity, 17, 1–14.

Tassetto, M., Kunitomi, M., Whitfield, Z.J., Dolan, P.T., Sánchez-Vargas, I., Garcia-Knight, M., *et al.* (2019). Control of RNA viruses in mosquito cells through the acquisition of vDNA and endogenous viral elements. *Elife*, 8, e41244.

Tehel, A., Brown, M.J.F. & Paxton, R.J. (2016). Impact of managed honey bee viruses on wild bees. *Curr. Opin. Virol.*

Teninges, D., Ohanessian, A., Christine, R.-M. & Contamine, D. (1979). Isolation and

Biological Properties of Drosophila X Virus. *J. Gen. Virol.*, 42, 241–254.

Thekke-Veetil, T., Lagos-Kutz, D., McCoppin, N.K., Hartman, G.L., Ju, H.-K., Lim, H.-S., *et al.* (2020). Soybean Thrips (Thysanoptera: Thripidae) Harbor Highly Diverse Populations of Arthropod, Fungal and Plant Viruses. *Viruses*, 12.

Thomas-Orillard, M. (1984). Modifications of Mean Ovariole Number, Fresh Weight of Adult Females and Developmental Time in DROSOPHILA MELANOGASTER Induced by Drosophila C Virus. *Genetics*, 107, 635–44.

Tijssen, P., Pénzes, J.J., Yu, Q., Pham, H.T. & Bergoin, M. (2016). Diversity of small, single-stranded DNA viruses of invertebrates and their chaotic evolutionary past. *J. Invertebr. Pathol.*, 140, 83–96.

Tsetsarkin, K.A. & Weaver, S.C. (2011). Sequential adaptive mutations enhance efficient vector switching by chikungunya virus and its epidemic emergence. *PLoS Pathog.*, 7, 1002412.

Unkless, R.L. (2011). A DNA virus of Drosophila. *PLoS One*.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M., *et al.* (2012). Primer3-new capabilities and interfaces. *Nucleic Acids Res.*, 40, 1–12.

Urayama, S.-I., Takaki, Y. & Nunoura, T. (2016). FLDS: A Comprehensive dsRNA Sequencing Method for Intracellular RNA Virus Surveillance. *Microbes Environ.*, 31, 33–40.

Vazeille, M., Gaborit, P., Mousson, L., Girod, R. & Failloux, A.-B. (2016). Competitive advantage of a dengue 4 virus when co-infecting the mosquito Aedes aegypti with a dengue 1 virus. *BMC Infect. Dis.*, 16.

Veech, J.A. (2013). A probabilistic model for analysing species co-occurrence. *Glob. Ecol. Biogeogr.*, 22, 252–260.

Viljakainen, L. & Jurvansuu, J. (2020). Discovery and Analysis of RNA Viruses in Insects BT - Immunity in Insects. In: (eds. Sandrelli, F. & Tettamanti, G.). Springer US, New York, NY, pp. 191–200.

Walker, P.J., Siddell, S.G., Lefkowitz, E.J., Mushegian, A.R., Adriaenssens, E.M., Dempsey, D.M., *et al.* (2020). Changes to virus taxonomy and the Statutes ratified by the International Committee on Taxonomy of Viruses (2020). *Arch. Virol.*, 165, 2737–2748.

Wallace, M.A., Coffman, K.A., Gilbert, C., Ravindran, S., Albery, G.F., Abbott, J., *et al.* (2020). The discovery, distribution and diversity of DNA viruses associated with Drosophila melanogaster in Europe. *bioRxiv*, 2020.10.16.342956.

Wallace, R.M., Gilbert, A., Slate, D., Chipman, R., Singh, A., Wedd, C., *et al.* (2014). Right place, wrong species: a 20-year review of rabies virus cross species transmission among terrestrial mammals in the United States. *PLoS One*, 9, e107539–e107539.

Wang, X.-H., Aliyari, R., Li, W.-X., Li, H.-W., Kim, K., Carthew, R., *et al.* (2006). RNA interference directs innate immunity against viruses in adult Drosophila. *Science*, 312, 452–454.

Wang, Y., Kapun, M., Waidele, L., Kuenzel, S., Bergland, A.O. & Staubach, F. (2020). Common structuring principles of the Drosophila melanogaster microbiome on a continental scale and between host and substrate. *Environ. Microbiol. Rep.*, 12, 220–228.

Wayne, M.L., Contamine, D. & Kreitman, M. (1996). Molecular population genetics of ref(2)P, a locus which confers viral resistance in Drosophila. *Mol. Biol. Evol.*, 13, 191–199.

Webster, C.L., Longdon, B., Lewis, S.H. & Obbard, D.J. (2016). Twenty-five new viruses associated with the drosophilidae (Diptera). *Evol. Bioinforma.*, 12, 13–25.

Webster, C.L., Waldron, F.M., Robertson, S., Crowson, D., Ferrari, G., Quintana, J.F., *et al.* (2015). The discovery, distribution, and evolution of viruses associated with drosophila melanogaster. *PLoS Biol.*, 13, e1002210.

Wickham, H. (2009). Elegant Graphics for Data Analysis. *Media*, 35, 211.

Wilfert, L. & Jiggins, F.M. (2013). The Dynamics of Reciprocal Selective Sweeps of Host Resistance and a Parasite Counter-Adaptation in Drosophila. *Evolution (N. Y).*, 67, 761–773.

Wilfert, L., Long, G., Leggett, H.C., Schmid-Hempel, P., Butlin, R., Martin, S.J.M., *et al.* (2016). Deformed wing virus is a recent global epidemic in honeybees driven by Varroa mites. *Science (80-.).*, 351, 594 LP – 597.

Wilkins, D. & Kurtz, Z. (2020). gggenes: Draw Gene Arrow Maps in “ggplot2.”

Williams, T., Bergoin, M. & van Oers, M.M. (2017). Diversity of large DNA viruses of invertebrates. *J. Invertebr. Pathol.*, 147, 4–22.

Woolhouse, M.E.J. & Gowtage-Sequeria, S. (2005). Host range and emerging and reemerging pathogens. *Emerg. Infect. Dis.*, 11, 1842–1847.

Woolhouse, M.E.J., Webster, J.P., Domingo, E., Charlesworth, B. & Levin, B.R. (2002). Biological and biomedical implications of the co-evolution of pathogens and their hosts. *Nat. Genet.*

Worobey, M., Pekar, J., Larsen, B.B., Nelson, M.I., Hill, V., Joy, J.B., *et al.* (2020). The emergence of SARS-CoV-2 in Europe and North America. *Science (80-.).*, 370, 564–570.

Wu, Q., Luo, Y., Lu, R., Lau, N., Lai, E.C., Li, W.X., *et al.* (2010). Virus discovery by deep sequencing and assembly of virus-derived small silencing RNAs. *Proc. Natl. Acad. Sci. U. S. A.*, 107, 1606–1611.

Xu, J. & Cherry, S. (2014). Viruses and antiviral immunity in Drosophila. *Dev. Comp. Immunol.*

Yampolsky, L.Y., Webb, C.T., Shabalina, S.A. & Kondrashov, A.S. (1999). Rapid accumulation of a vertically transmitted parasite triggered by relaxation of natural selection among hosts. *Evol. Ecol. Res.*, 1, 581–589.

Yu, G. (2020). Using ggtree to Visualize Data on Tree-Like Structures. *Curr. Protoc. Bioinforma.*, 69, e96.

Zhang, K.-Y., Gao, Y.-Z., Du, M.-Z., Liu, S., Dong, C. & Guo, F.-B. (2019a). Vgas: A Viral Genome Annotation System . *Front. Microbiol.* .

Zhang, Y.-Z., Chen, Y.-M., Wang, W., Qin, X.-C. & Holmes, E.C. (2019b). Expanding the RNA Virosphere by Unbiased Metagenomics. *Annu. Rev. Virol.*, 6, 119–139.

7 Appendix

Site	Short code	Latitude	Longitude	Elevation (m)	Trap location
River almond walkway (Crammond)	CR	55.975472	-3.301978	21	Nature trail passing along a clifftop through a mixture of woodlands, marsh, ponds and meadowland.
Corstorphine hill Local Nature Reserve	CS	55.953478	-3.276433	136	Woodland near the walled gardens surrounding Clerwood house and Corstorphine hill
Sighthill public park	SI	55.92865	-3.285806	49	Wooded area at corner of sighthill public park. The park is bounded by community woodlands.
Bangholm playing fields	TR	55.974219	-3.202067	34	Trees at the edge of Trinity Academy Playing fields (Bangholm).
Edinburgh Botanic gardens	BG	55.965681	-3.205347	18	Woodland around west entrance gate at royal botanic gardens near Inverleith park
Hermitage of Braid local nature reserve	BR	55.918911	-3.210475	94	Trees near the road and coffee shop within the hermitage of braid area
Hillend country park	HL	55.885692	-3.201314	189	Woodland on the site of Hillend country park, near to the ski centre
Centre for Ecology and Hydrology - Bush estate	CE	55.861661	-3.204533	195	Woodland opposite the centre for ecology and hydrology at the bush estate.
Inveresk lodge gardens	IN	55.935267	-3.044133	25	Area close to fruit trees in the gardens of Inveresk Lodge in Musselburgh
Gilmerton shopping centre carpark (near Edinburgh insect and butterfly world)	IB	55.895978	-3.100689	76	Trees across the carpark from the insect and butterfly world buildings
Lasswade riding centre	LW	55.877925	-3.123814	78	Trees in woods surrounding Edinburgh & Lasswade riding centre
Cockenzie house gardens	CK	55.970003	-2.962939	7	Trees on outskirts of Cockenzie house and gardens
Oxenfoord castle grounds	OX	55.879017	-2.980708	145	Trees particularly close to road alongside Oxenfoord castle, and surrounding estate
Vogrie country park	VG	55.857519	-3.000206	172	Woodland around car park at Vogrie country park
Gosford farm	GF	56.003525	-2.871544	11	Woodland to the back of Gosford bothy farm shop, part of Gosford farm
Elvingston stud & Riding centre	EL	55.958794	-2.864553	87	Trees around Elvingston stud riding school and horse breeding centre, close to Elvingston house
East coast organics farm (Haddington)	EC	55.924986	-2.875692	107	Trees around east coast organics farm shop, near Haddington
Tranent parish church graveyard	TN	55.949656	-2.957928	72	Overgrown, woody area at edge of Tranent parish church graveyard
Dalkeith country park	DK	55.899067	-3.065975	51	Wooded area around Dalkeith country park and adventure centre, near the coffee shop
Burdiehouse burn valley park Local Nature Reserve	BN	55.910253	-3.147881	95	Wooded area in local nature reserve, access from Ellen's Glen Loan

Table S2 1 Sampling sites used to collect Drosophila. The table details the location, and positioning of banana and yeast baited traps used to catch Drosophila in 2016-2018. I always got proper permission to place traps and collect flies from the landowner, business owner, or local council manager.

Virus / Virus segment	Forward primer name	Forward primer sequence	Tm	Reverse primer name	Reverse primer sequence	Tm	Product size	Notes
Inveresk virus	Inv_3077_F	GATGTGTTTGACATTGAGTACATG	56.6	Inv_3964_R	GCACGATATGACATAGGACC	55.5	887	All primer combinations work
	Inv_3322_F	AACGATGAGGAAAGTATCATGG	56	Inv_4416_R	GGATTCCTGATTTCATCAATGAAC	56.4	1094	
Sunshine virus	DrosMix_Bunyavirus_L4_00346F	CAAAGAAGCAAGAGTCCATCC	56.9	DrosMix_Bunyavirus_L4_01828R	GCTAATTTCTCTATTGCCCCC	56.4	1482	Only one working primer pair
Sighthill Virus - L segment	DrosMix_Bunya_Chizelike_02051F	GGATCCTTCTCATTAACAGCC	56.1	DrosMix_Bunya_Chizelike_02368R	GGTACTTCTTACTAACGCTCC	55.6	317	Only one working primer pair
Sighthill Virus -M segment	Bunya_H_Tick_2like_01064F	CTCCTGAATGTTTCTCCAAGG	55.7	Bunya_H_Tick_2like_01330R	GAGACAAGAGGATGAAAGTGG	56	266	Single F primer works with either R primer
				Bunya_H_Tick_2like_01668R	GTCCTCAACTCCATTAATCCC	56.2	604	
Vogrie Virus - segment 1	DrosMix_Reovirus1_01808F	CGACCCATATACATCAACCC	55.4	DrosMix_Reovirus1_02617R	GATGGGTCTCGAAATATCATCC	56.2	809	Only one working primer pair
Vogrie Virus - segment 5	DrosMix_Reovirus2_01248F	ATGCCGGATCAATTCAGG	54.7	DrosMix_Reovirus2_01769R	GTATGGGATATTCGCCACC	55.2	521	Only one working primer pair
Vogrie Virus - segment 2	Vog_S2_813_F	GAAACTATCTCGGTCCACAAG	56.1	Vog_S2_1528_R	AGAACGCATATAAGCATCCATC	56.5	715	Only one working primer pair
Tranent Virus - L segment	Tran_L_2340_F	GACTCAATGACAGTGAACCATTG	58	Tran_L_3099_R	TGACCCATTGAACAACTCTCTC	58	759	All combinations of these 4 primers work
	Tran_L_2736_F	AGCAACATGGGAAAGGAATATG	56.7	Tran_L_3451_R	ACACCTTGCATTGATTCATTCC	57.5	715	
Tranent Virus - M segment	Tran_V_M_1004_F	GAGGATGAACTTGTGAAAGGG	56.5	Tran_V_M_1533_R	GAGTTGCATTGCATTTTGCA	56.4	529	All combinations of these 4 primers work
	Tran_V_M_1041_F	AGATGTCAAGCTGAACCTCTT	56.9	Tran_V_M_1689_R	GCTGAAGAAAGATCCAACCTC	56.6	648	
Tranent Virus - S segment	Tran_V_S_707_F	TTGAATCCAAGATGGCAGAGG	58	Tran_V_S_1355_R	GTAGAAAGTGCTCCCCAATCTT	58	NA	941F works with either R, 707F only with 1570R
	Tran_V_S_941_F	GAGCACTGATTGAACGAACATG	58.3	Tran_V_S_1570_R	TCCACCCAGAGAAATCAACTTC	58	629	
Lasswade Virus	DrosMix_Vesiculovirus1_00565F	CAGTTGTATCCGTGGATCC	55.1	DrosMix_Vesiculovirus1_01662R	CTCTACTTGAAGAGAAGCGG	55.3	1097	Only one working primer pair
Cockenzie Virus	DrosMix_Cripa_DCV-like_02067F	TCACTCGCTAGCAATTCG	56.6	DrosMix_Cripa_DCV-like_02720R	TATCCTGTTCTCCTGTTAAGGG	56.3	653	Only one working primer pair
Burdiehouse burn Virus	DrosMix_Chuvirus_02165F	ATTCAGAACTCTCCTCTGATCC	56.1	DrosMix_Chuvirus_02866R	CTGACAAGCCTTCTATGATGG	56.6	701	Either F primer works with the single R primer
	DrosMix_Chuvirus_02283F	GAGCACTGTATTGAAACTCCC	55.5				583	
Dalkeith Virus	Shy1_Chuv_Lgene_3105_F	CAAGCTTGTTATTATGGCCTCAG	57.8	Shy1_Chuv_Lgene_3834_R	CCTGTACATCACGGTTCCATG	58.7	729	3105F works with either R, 3169F only with 3834R
	Dalk_V_3169_F	TTGTCAGAGGAGAGAATGATATGC	58.1	Dalk_V_4029_R	CCAACCAGGGAGCTGCA	59.2	860	
	Cramm_V_4951_F	TTCGCTGACTTTCTGATGAAGA	57.7	Cramm_V_5856_R	GACTCCATGAGAAGCCCAAT	57.3	905	

Crammond Virus	Cramm_V_5273_F	TGGTACAAGATATCATGTCGGA	56.4	Cramm_V_5995_R	ACTGGTCTGCTAATCGTTTCAT	57.8	722	All combinations of these 4 primers work
River almond virus	River_almond_6137_ F	CACGTGTGGTTAACTGTTGTCC	60	River_almond_7054_R	GCATACCGGTGAAAGAGTCG	58.7	918	All primer combinations work apart from 6220F w 7103R
	River_almond_6220_F	GATCTATCACTTGACGCGTGG	58.9	River_almond_7103_R	AGGAAAGTAGCACTATCGCCAC	60.2	884	
Hillwood park virus	Hillwood_park_5499_F	GTGTTTCTTTACCGATGATCTCAC	57.9	Hillwood_park_6257_ R	ATGCAGCGTAATTGTCTGTTGC	60.4	758	Only one working primer pair
Craighall virus	Craighall_3735_F	GGCAGGACTTTTCATGTGGTG	59.7	Craighall_4429_R	CCAACCCATCACTAGCAAAAGC	60.1	695	All primer combinations work except 3735 w 4429
	Craighall_3771_F	AGAAGTGTGTTTGGCCATGTTG	59.9	Craighall_4547_R	TCTGTGAAGTCGGCTGATAACC	60.1	777	
Inverleith virus	Inverleith_4714_F	GTCTGTGATGGTATTGTTGCCG	59.9	Inverleith_5560_R	AATTCCAGAAGTAGAGCCCAGC	60.1	847	All primer combinations work
	Inverleith_4767_F	GCTTGATGATTGGTGGAGTATGC	60	Inverleith_5724_R	AACGCCGGGCTGCTTATAC	60.5	598	
Midmar virus - segment 1	Midmar_S1_1228_F	CTGGACAAATGACGATCGCC	59.3	Midmar_S1_1946_R	CCACTTCATCTGGTTGCACATC	59.8	719	All primer combinations work
	Midmar_S1_1320_F	CGGAGTGGATTGCAACTATGG	59.1	Midmar_S1_2007_R	GTAGGATCTGAACCACACTCTTC	58.3	688	
Midmar virus - segment 2	Midmar_S2_951F	GGATGGATTGTGAGGGAAAGGG	60.7	Midmar_S2_1562R	CTCATCGCAACCGTAACGTTG	60.2	633	All primer combinations work
	Midmar_S2_1011F	AGGTCGTGTCTTCATCTTCTGG	59.8	Midmar_S2_1834R	TCTCCAGGTGCTGCAATGC	60.7	822	
Glencorse burn virus - segment 1	GlencorseB_S1_1249F	TGACTACATGTATAAGCGGTTGG	58.3	GlencorseB_S1_1848R	CCTCTCCAAAAGCGTATTCTACC	58.9	600	All primer combinations work
	GlencorseB_S1_1293F	GGAAAAGTACTTCGTCTGGAATGG	59.6	GlencorseB_S1_2025R	CCACTAATTCGGAAACTCCATGC	59.9	756	
Glencorse burn virus - segment 2	GlencorseB_S2_370F	GTTTGTGGCCCTGAACTGG	59	GlencorseB_S2_1032R	CTTCCTCAGAAACGGCAAAAGG	60	663	All primer combinations work
	GlencorseB_S2_412F	ACAAGTAAGTGTTGCGAGATGG	58.9	GlencorseB_1049R	AATTCCCGCTTTCCTCTCTTCC	60.4	638	
Glencorse burn virus - segment 3	GlencorseB_S3_1063F	GGAATCCCATATCGCAAACGG	59.5	GlencorseB_S3_1669R	TACGTTCTCCATTTTGCAAGCC	59.8	607	All primer combinations work, but 1092F with either R primer creates less secondary product.
	GlencorseB_S3_1092F	AAGACAGCTGCGTTAAATGTGG	59.8	GlencorseB_S3_1687R	AGAGCCATAACCCCTTGATACG	59.6	596	
Glencorse burn virus - segment 4	GlencorseB_S4_13F	TTTCGTATAATGGTTCGCAGCG	59.7	GlencorseB_S4_681R	CGGGTCATGTGCCTCTATTGG	60.8	669	All primer combinations work, but 13F w 976R produces weak bands
	GlencorseB_S4_138F	TGCAGAAACGGGCACTTGG	61.2	GlencorseB_S4_976R	TCTCTCCACTATGTTTGCGACC	60.1	857	
Glencorse burn virus - segment 5	GlencorseB_S5_408F	CCAGGGATTAATGTGGCATTCG	59.7	GlencorseB_S5_1019R	GCTCATATGAGGTAGAGCATGACG	60.9	637	All primer combinations work
	GlencorseB_S5_445F	ACGCTTCCGATTTAGTTTCATGG	59.6	GlencorseB_S5_1044R	GCGAGAGTTTTTGTACGTTCTCC	60.1	601	

North esk virus - L segment	North_esk_Lseg_3049F	GCTGATGTCAATGTCAAGAGGC	59.9	North_esk_Lseg_3744R	GTAACCACCAAGATCAACTGGC	59.5	696	All primer combinations work
	North_esk_Lseg_3193F	GTTCGAAGTGTCATTGGATGGC	60.2	North_esk_Lseg_3813R	CTGAACCTGGTGTATGTTTCAGC	59.8	643	
North esk virus - M segment	North_esk_Mseg_1129F	AGATGGAGGTGCATGCTGG	59.8	North_esk_Mseg_1860R	GTTGGTATGTCAATCTCGCACC	59.6	732	All primer combinations work
	North_esk_Mseg_1245R	GTTGTGATACATGGTGCTGGC	59.9	North_esk_Mseg_2199R	AATGTGGATGAGAGGGAAGCC	59.8	955	
North esk virus - S segment	North_esk_Sseg_371F	AGGGATCACAAGCAGAGTGG	59.4	North_esk_Sseg_1197R	CCTCTCGAACAAATCTTGATCCC	59.1	827	All primer combinations work except 371F w 1310R
	North_esk_Sseg_487F	TCAGGTACGCCATTCGATTGG	60.5	North_esk_Sseg_1310R	GCATCCTTCCTCGACTGCTC	60.5	824	
Gosford virus	Gosford_1012F	TCGATAGGATTATTACGGCCCG	59.6	Gosford_1742R	TGGTACCACAGCACATTCCTC	60	731	All primer combinations work
	Gosford_1126F	TGAACGTGGCAGCACATATG	58.9	Gosford_1786R	CCTCGATACTTAATGCGTGACG	59.3	661	

Table S2 2. RT-PCR Primer assays designed for newly described RNA viruses.

Virus	Virus Classification	First Description	Genbank Accession No.	Forward primer name	Forward primer sequence (5'-3')	Reverse primer name	Reverse primer sequence (5'-3')	Product length (bp)	Tm (ºC)	Extension time	Primer source
Chaq satellite	Unclassified	(Webster et *al*., 2015)	KP714088	222_140F	AACAGAACGWCTGCTTTTTGGAAATCC	222_660R	TCCATGTCCTGTHGGGTCTATCTG	520	60	45	(Webster et *al*., 2015)
				Chaq_514F_short	CGAAGTAACATACCAGCCATGG	222_660R	TCCATGTCCTGTHGGGTCTATCTG	126	57	60	Megan Wallace (Chapter 4)
Cherry Gardens virus	*Rhabdoviridae* (-ssRNA)	(Webster et *al*., 2016)	KU754524	Dsub_CherryGardens_03600F	CTGGTGATGGAACTGGTGGAG	Dsub_CherryGardens_04416R	TGGTCGTGGAGCCTGATTAG	814	59	60	Nathan Medd (Obbard lab) thesis
Craigie's Hill virus	Alphanodavirus (+ssRNA Segmented)	(Webster et *al*., 2015)	KP714084 KP714085	NV-L_1.69F	GCAAAATCCGTGGTTCATACCAG	NV-L_2.89R	CTTAACAGGACGCTCCAAGTGGAT	1200	60	90	(Webster et *al*., 2015)
				674_100F	CCTATCTGTCAAGCTGTWCTGCCAAC	674_1540R	GTGTGCCAACTAGGCTCAGGAG	1440	62	90	
Dimm Nora virus	Picornavirales (+ssRNA)	(van Mierlo et *al*., 2014)	KF242511	Dimm_Nora_A_10358F	TGAATCCTTGGATGCGAACGG	Dimm_Nora_A_11173R	TTGATGCTGCTCCTAACACGG	816	61	60	Nathan Medd (Obbard lab) thesis
				Dimm_Nora_B_03392F	AGGCATTACACGCTCGATTCC	Dimm_Nora_B_04071R	TTCACTCGCCATAGGAACACC	680	61	60	
Dimm Sigma virus (L & N genes)	Sigmavirus (Rhabdoviridae) (-ssRNA)	(Longdon et al., 2011)	KR822814	DImmSV_L_F2	TGATAGATCCTTCCGCCATC	DImmSV_L_R2	GAATGCTCCAACCCTTTTGA	696	56	60	Longdon et al. 2010
				DImmSV_N_ F1	CTAGCATTTGCCGGGGATAAA	DImmSV_N_R1	AATGCATTCCTGGTCTTTGG	782	56	60	
Dsub Nora virus	Picornavirales (+ssRNA)	(van Mierlo et *al*., 2014)	KF242510.1	Dsub Nora 1665F	GCTTACGAGAAAGCAGAACG	Dsub Nora 2415R	ACAGTACTCTCCCAAATACCTTG	787	57	60	Designed for this thesis - Megan Wallace
Galbut virus	Unclassified partitivirus	(Webster et *al*., 2015)	KP714100 KP714099	407_170F	GATCGAGATGGAACTCCRCTCTC	407_750R	GCCKCATACTTGGTGCTGCCAACTG	580	60	60	(Webster et *al*., 2015)
Muthill virus	cf. Negevirus (+ssRNA)	(Webster et *al*., 2016)	KU754517	Dsuz_Muthill_06921F	AATGCGTTTCCTAGCCCACAC	Dsuz_Muthill_07613R	ATGGCGCGTTGTTTGTTTCG	693	61	60	Nathan Medd (Obbard lab) thesis
				Muthill_6795F_sh	GCGTCACAAGTCCAAGTTCATG	Muthill_6943R_sh	GTGTGTGGGCTAGGAAACGC	148	61	60	Megan Wallace (chapter 4)
Pow burn virus	Picornaviridae (+ssRNA)	(Webster et *al*., 2016)	KU754519	Dobs_PowBurn_B_03155F	CACAAGAAGAAGCGTGACTC	Dobs_PowBurn_B_03926R	TGTTAGCCTCTCCGTATGC	772	55.5	60	Nathan Medd (Obbard lab) thesis

Corseley	Unclassified (+ssRNA)	(Webster et *al*., 2016)	KU754520	DsubDsuz_Corseley_00511F	ACGTGTTGAGCGAGGAGTAC	DsubDsuz_Corseley_01311R	TTCTGCTACACTCATGCTGGC	801	60	60	Nathan Medd (Obbard lab) thesis
Craigmillar park virus	Alphanodavirus (+ssRNA Segmented)	(Webster et *al*., 2016)	KU754525 KU754526	Dsus_Craigmillar_seg1_01679F	AGCAGTATCCGTGGTTCATCC	Dsus_Craigmillar_seg1_02483R	TCATGTTGAGGCCAGGAATCAG	805	59	60	Nathan Medd (Obbard lab) thesis
				Dsus_Craigmillar_seg2_00112F	GTGCCGACTAAGGTTGCTCTC	Dsus_Craigmillar_seg2_00806R	CGTGTGAGTTGATTCCACAGAC	695	59	60	Nathan Medd (Obbard lab) thesis
Grom virus	cf. Sobemoviruses and Poleroviruses (+ssRNA)	(Webster et *al*., 2016)	KU754506	Dobs_Grom_00728F	CGGCAGGTCACAACATCTCAT	Dobs_Grom_01554R	GCTTTGGGTGACTGTGGACT	826	60	60	Nathan Medd (Obbard lab) thesis
La Jolla virus	Iflavirus (+ssRNA)	(Webster et *al*., 2015)	KP714073	SB-L_6.77_F	GTGGAGTAAAGCAACGACTTGG	SB_ASSAY_1R	CAACTGCRTGTTTGAGTTCCCAACGA	1300	60	90	(Webster et al., 2015)
Lye green virus	Rhabdoviridae (-ssRNA)	(Webster et *al*., 2016)	KU754522	Dobs_Lye_Green_04753F	TTGTCGAGAATAGCAGGAGTCC	Dobs_Lye_Green_05516R	AGTCCGGTCCTAGTCTGAAGC	764	59	60	Nathan Medd (Obbard lab) thesis
Motts mill virus	cf. Sobemoviruses and Poleroviruses (+ssRNA)	(Webster et *al*., 2015)	KP714076 KP714077	Luteo_1F_364	AATAAATCATAAATCGTGCTTGTTTCCTTGGC	Luteo_2R_1758	AATAAATCATAAGGTTGAACCAGTCGGTGAAT	1400	67	60	(Webster et al., 2015)
				221_100F	ACAGCAGAGTTCTTGCGAGGAGC	221_795R	GACTGCCACGTCTCATGCTTCACTTC	695	65.5	60	
Dmel Nora virus	Picornavirales (+ssRNA)	(Habayeb et *al*., 2006)	NC_007919	Nora_6220F	GACCATTGGCACAAATCACCATTTG	Nora_7210R	TCTTAGGCCGGTTGTCTTCACCC	990	60	60	(Webster et al., 2015)
				Nora_qPCR_3F_flap	AATAAATCATAAGGTGTAGCAGGTCGTATTCTGC	Nora_qPCR_3R_flap	AATAAATCATAACAATGGCTGAAACTGCTGTTCCTGC	120	60	60	(Webster et al., 2015)
Prestney burn virus	cf. Sobemoviruses and Poleroviruses (+ssRNA)	(Webster et *al*., 2016)	KU754507	Dsub_Prestney_Burn_A_00190F	GGCCAATTGACTGAATCGGACC	Dsub_Prestney_Burn_A_00880R	TGGTGGTGGTTGGTTGATCG	690	61	60	Nathan Medd (Obbard lab) thesis
				PrestneyB_00737F_short	CGGCAGGTCACAACATAATATCAG	Dsub_Prestney_Burn_A_00880R	TGGTGGTGGTTGGTTGATCG	143	60		Megan Wallace (chapter 4)
***Drosophila* A virus**	cf. Permutotetraviridae (+ssRNA)	(Plus et *al.*, 1975)	NC_012958	DAV_3300F	TGCAAGTAAGCTCTTGCCAACCCT	DAV_3940R	AGATACCACTTACGGGTGGTTGC	640	60	60	(Webster et al., 2015)
Kinkell virus	Iflavirus (+ssRNA)	(Webster et *al*., 2016)	KU754510	Dsus_Kinkell_03036F	TGTTGTGTACGAGCTGTGGTC	Dsus_Kinkell_03777R	ACACCGATGAGAGCGAGGATG	741	59	60	Nathan Medd (Obbard lab) thesis
Buckhurst virus	cf. Negevirus (+ssRNA)	(Webster et *al*., 2016)	KU754516	DobsDsub_Buckhurst_06402F	TTCGGAGGTGGTGATAACGC	DobsDsub_Buckhurst_07190R	AGCAGCTTGTGTATCCAACC	738	58	60	Nathan Medd (Obbard lab) thesis
Withyham virus	Rhabdoviridae (-ssRNA)	(Webster et *al*., 2016)	KU754523	Dobs_Withyham_00425F	CTGGGCACTTTGGAATCCTC	Dobs_Withyham_00970 R	ATGTGTCCGCCATCATCAAC	545	57.5	60	Nathan Medd (Obbard lab) thesis

Larkfield virus	Totiviridae (dsRNA)	(Medd et *al.*, 2018)	MF893249	DrosMix_Totivirus_01124F	GTACCAGATCATTGCTATGACC	DrosMix_Totivirus_02065R	GGAATCGTGTATATCGAAGAGC	941	61	60	Medd et al., 2018
Thika virus	Picornavirales (+ssRNA)	(Webster et *al.*, 2015)	KP714072	Cripa_C3_5625F	CTTCGAAGCATCYCTGCATCGTAAAG	Cripa_univ_6560R	GCACCCACAGCTAGCATRTCTGG	900	60	60	(Webster et al., 2015)
***Drosophila* C virus**	Cripavirus (Dicistroviridae) (+ssRNA)	(Jousset et *al.*, 1972)	NC_001834	DCV7	AGTATGATTTTGATGCAGTTGAATCTC	DCV8	GAAGCACGATACTTCTTCCAAACC	524	59.5	60	(Kapun et al., 2010)
***Drosophila melanogaster* American Nodavirus**	Nodavirus (+ssRNA)	(Wu et *al.*, 2009)	GQ342966 GQ342965	Dmel_ANV_2_00373F	CACCTGGAGTCGTTATCTGG	Dmel_ANV_2_01021R	GGCCGAATTGATACAGCATAGC	695	58	60	Nathan Medd (Obbard lab) thesis?
Drosophila X virus	Entomobirnavirus (dsRNA)	(Teninges et *al.*, 1979)	NC_004177 NC_004169	DXV_SegB_F	CATGCCAATCAGTGATGACG	DXV_SegB_R	GCTGTTGTCATTGCCACATC	738	57.7	60	Nathan Medd (Obbard lab) thesis
Charvil virus	Flaviviridae	(Webster et *al.*, 2015)	KP714089	Dmel_Charvil_01343F	AGGAGTTGACGGATGATGAGG	Dmel_Charvil_01868R	GTTGGTGCGTACTTGTGAACC	525	59	60	Nathan Medd (Obbard lab) thesis?
Kallithea virus	Nudivirus (dsDNA)	(Webster et *al.*, 2015)	KP714101-KP714108	NudiPif1_F	CGACATCACATTCGACCCATATCC	NudiPif1_R	TCCCATAAAGTGCGATCCCATAG	970	62	90	(Webster et *al.*, 2015)
Kilifi virus	Picornavirales (+ssRNA)	(Webster et *al.*, 2015)	KP714071	Dmel_Kilifi_A_03173F	CTGGTCGCTGTCATGAGGTTG	Dmel_Kilifi_A_03875R	ATTGGACGCTGTGATGTCGC	703	60	60	Nathan Medd (Obbard lab) thesis?
Hermitage virus	Unclassified (RNA)	(Webster et *al.*, 2016)	KU754511 KU754512	Dimm_Hermitage_A_00082F	ACATGTATCAACCACCGCGAC	Dimm_Hermitage_A_00845R	ACCGATTTGACACCAGGCTTG	763	60	60	Nathan Medd (Obbard lab) thesis
				Dimm_Hermitage_B_01531F	TTGAAGGTGACGGAAGCCATC	Dimm_Hermitage_B_02182R	CGTTCTTGGTGTGCTCATCG	651	59	60	Nathan Medd (Obbard lab) thesis
Twyford virus	Iflavirividae (+ssRNA)	(Webster et *al.*, 2015)	KP714075	SB-L_1.51_F	CGCAGTCAGTTTGCATCAGG	SB.TWY_2.57_R	CTCAGCTAAGGAGCCTTCCAT	900	62	90	(Webster et al., 2015)
Tartou virus	Unclassified (+ssRNA)	(Webster et *al.*, 2016)	KU754521	Sdef_Tartou_00404F	CGCATTGAATACGCCAGAACC	Sdef_Tartou_01080R	GCTATCCGAGACATGTGTTGC	677	59	60	Nathan Medd (Obbard lab) thesis
Berkley virus	Picornavirales (+ssRNA)	(Webster et *al.*, 2015)	SRA SRR070416	Dmel_Berkeley_02000F	GGTAAGGCTGGATGCTTGGT	Dmel_Berkeley_03057R	AGGGAGACGCAAGCATTTGA	1058	60	90	Nathan Medd (Obbard lab) thesis
Brandeis virus	Unclassified cf. Negevirus and Virgaviridae (+ssRNA)	(Webster et *al.*, 2015)	SRA SRR486227	Dsuz_Brandeis_05938F	TCTTCAACCGCATGTCCGTG	Dsuz_Brandeis_06869R	GGTGAAGGTGGTGGCATGAC	932	60	90	Nathan Medd (Obbard lab) thesis
Versanto virus	Bidensovirus (ssDNA)	(Kapun et *al.*, 2018)	KX648533 KX648534	Dmel_Vesanto_02080F	ATTGCAGACGACGACACCAC	Dmel_Vesanto_02920R	CCTTGACACGCTTACCATGC	841	60	60	Nathan Medd (Obbard lab) thesis

Table S2 3 Known Drosophila virus PCR primers. Details of the virus primers used to assay both single and pooled Drosophila by species for host range, and prevalence of known Drosophila viruses.

Sequencing pool	Virus	Original description	Most recent assembly accession	Length (kbp)	Details	Updated Length (kbp)	Details
Dec '16 - Jun '17	***Drosophila* tristis sigma virus**	Longdon *et al.* (2015)	KR822818.1	5.4	A –ssRNA Rhabdovirus described from a pool of *D. tristis*, and submitted as the L gene only, encoding the partial RdRp (1782aa)	13.6	New assembly encodes a partial nucleocapsid protein (493aa), polymerase-associated protein (287aa), PP3 (340aa), matrix protein (227aa), glycoprotein (630aa), and RdRp (2138aa), representing the near-complete genome of this Rhabdovirus.
	Kinkell virus	Webster et al. (2016)	KU754510.1	6.8	A +ssRNA Iflavirus described from a pool of *D. subsilvestris*. The original assembly contained a 226 bp gap, and encoded a 2122aa putative polyprotein.	10.2	New assembly contains no gaps and completes the putative polyprotein (2967 aa). It also contains two other hypothetical proteins, one 45aa, and one 55aa.
	Larkfield virus	Medd *et al.* (2018)	MF893249.1	6	A dsRNA Totivirus described from a pool of *D. suzukii*, encoding 3 ORFs; two hypothetical proteins and a putative RdRp	6	New assembly covers two single bp gaps in the original assembly, removing them. It encodes three proteins, like the original assembly, a capsid protein (1024aa), a hypothetical second protein (53aa), and a putative RdRp (681aa).
Jul-Oct '17	**Lye green virus**	Webster *et al.* (2016)	KU754522.2	14.3	A -ssRNA Rhabdovirus encoding 5 proteins, and closely related to *D.busckii* Rhabdovirus. Initially described from *D.obscura*. The most recent assembly contained 336+423+377bp gaps in the RdRp.	14.2	New assembly encodes seven proteins; VP1 (putative nucleocapsid protein – 483aa), VP2 (putative polymerase-associated protein – 451aa), VP3 (putative matrix protein – 262aa), a putative glycoprotein (639aa), VP5 (101aa), an RdRp (2324aa), and VP7 (57aa) and contains no gaps.
	Tartou virus	Webster *et al.* (2015)	KU754521.1	1.8	+ssRNA virus genome, originally described in *S.deflexa*, encoding a single ORF. Closely related to *Diaphorina citri* associated C virus.	2.7	New assembly extends this virus genome, and has 96.4% similarity at the protein level to the original virus, It also only encodes a single protein, hypothetical protein 1 (466aa).

Apr-Jun '18	**Prestney burn virus**	Webster *et al*. (2016)	KU754507.1	3	A +ssRNA virus originally described in *D. subobscura*, closely related to Teise, Grom and Mott mill viruses (cf. Polerovirus/ sobemovirus). The original assembly had a partial sequence for the second ORF, with an estimated 198bp gap in the assembly.	3	New assembly fills in the gap of the previous assembly, but lacks the polyA tail. It encodes two proteins, protein 1 (587aa), and a putative RdRp (347aa, extended from 253 in the original assembly)
Jul-Oct '18	**Hermitage virus**	Webster *et al.* (2016)	KU754511.1 + KU754512.1	3.2 + 3.5	A putative dsRNA virus originally described from *D. immigrans.* Originally submitted as two separate contigs which encoded a putative polyprotein, and with some similarity to *Flaviviridae*.	13.2	New assembly encodes a single protein, a partial polyprotein (4165aa) with 53.19% protein identity to Takaungu virus (AMO03219.1), and 34.95% to *Lampyris noctiluca* flavivirus 1 (QBP37018.1) polyproteins.
	Withyham virus	Webster *et al.* (2016)	KU754523.1	6.9	A –ssRNA Rhabdovirus, closely related to *Dsubobscura* Rhabdovirus, and originally described from *D.obscura.* Originally submitted as a partial genome encoding a putative L protein gene, with an 83bp gap.	15.7	New assembly, likely completing the genome, encodes seven proteins; a putative nucleocapsid protein (490aa), putative protein 2 (399aa), putative protein 3 (300aa), putative protein 4 (104aa), putative glycoprotein (694aa), putative protein 6 (104aa) and putative RdRp (2255aa)

Table S2 4. Extended *Drosophila* virus genome assemblies. Details of the eight already published virus genomes extended by five rounds of metagenomic sequencing on mixed species pools of *Drosophila*. The currently published genome assemblies were either extended, or had gaps filled in using assemblies from contigs from one of my sequencing pools. I was unable to assign host range to any of these viruses in this chapter, due to the mixed species in the pools.

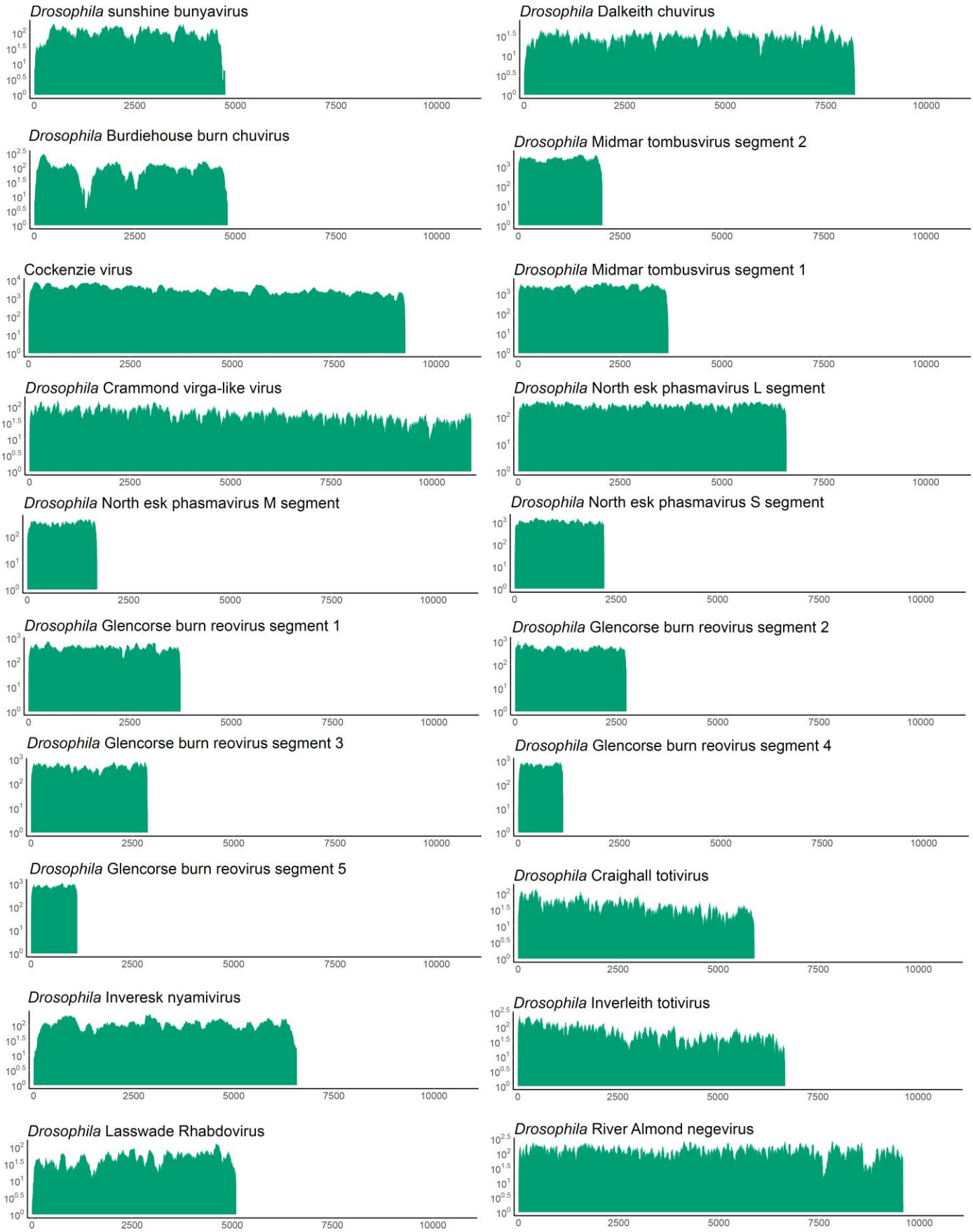
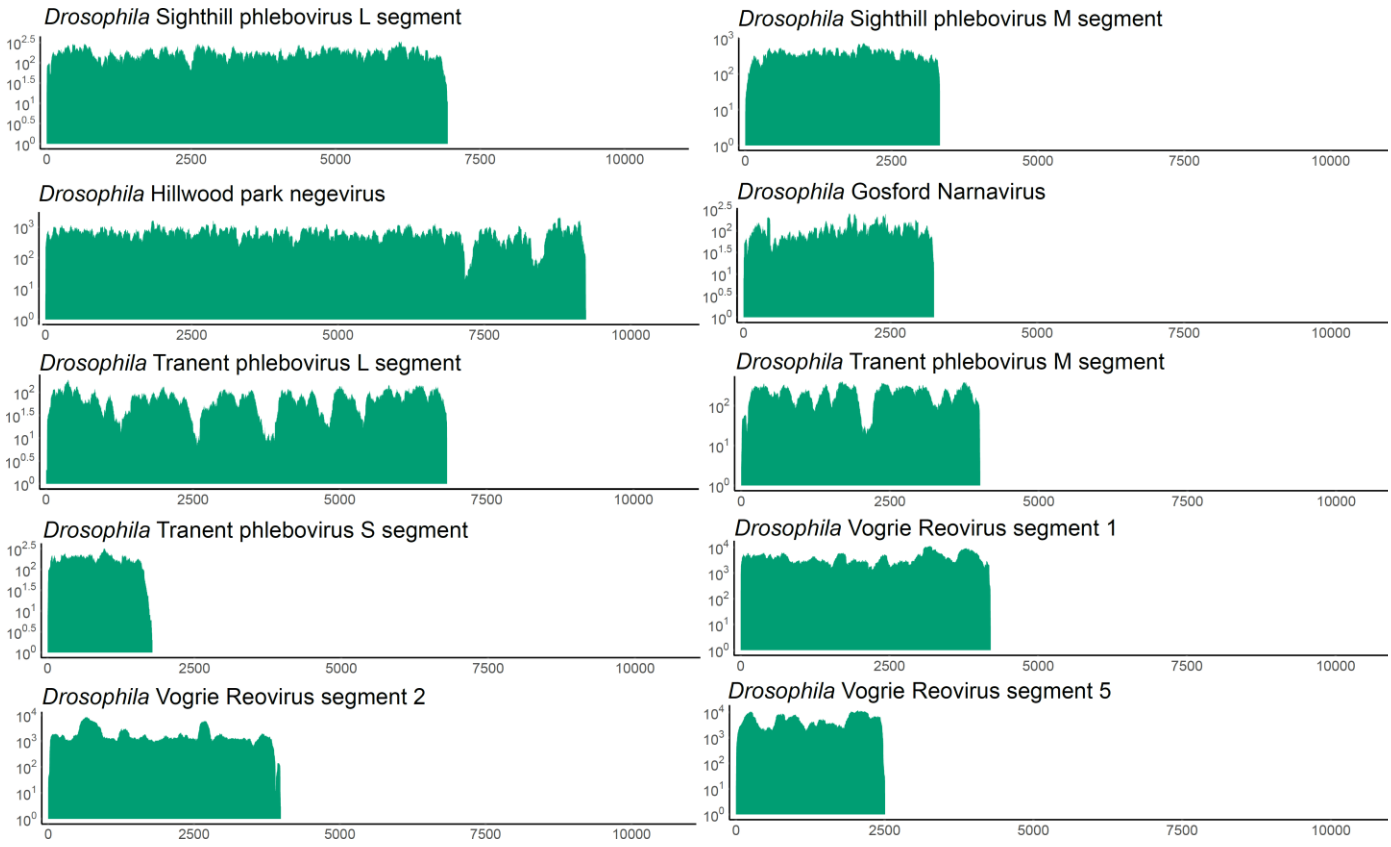


Fig. S2. 1 Read mapping against genomes of newly described Drosophila viruses (cont. on next page). The plot shows the depth of coverage when reads from all five datasets were mapped back against virus genomes using bowtie2 (sensitive).


Drosophila Sighthill phlebovirus L segment
Drosophila Sighthill phlebovirus M segment
Drosophila Hillwood park negevirus
Drosophila Gosford Narnavirus
Drosophila Tranent phlebovirus L segment
Drosophila Tranent phlebovirus M segment
Drosophila Tranent phlebovirus S segment
Drosophila Vogrie Reovirus segment 1
Drosophila Vogrie Reovirus segment 2
Drosophila Vogrie Reovirus segment 5
0
2500
5000
7500
10000

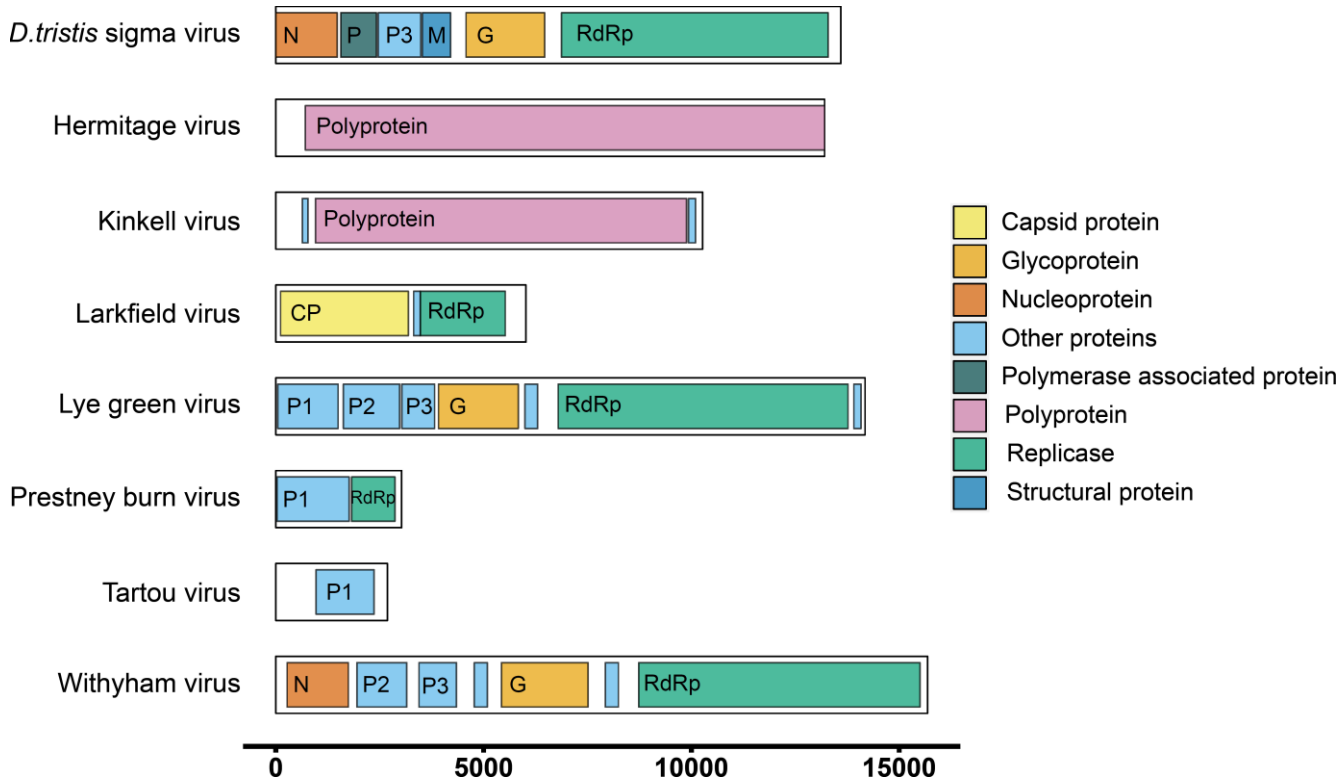


Fig. S2. 2 Genome organisation of extended Drosophila viruses. Genome organisation of eight extended RNA virus genomes, found in mixed species pools of Drosophila. All viral RNAs which exist as a negative sense strand are represented in their positive orientation. Colours indicate the hypothetical protein coding regions, and any functional annotations found by homology to other known virus proteins.

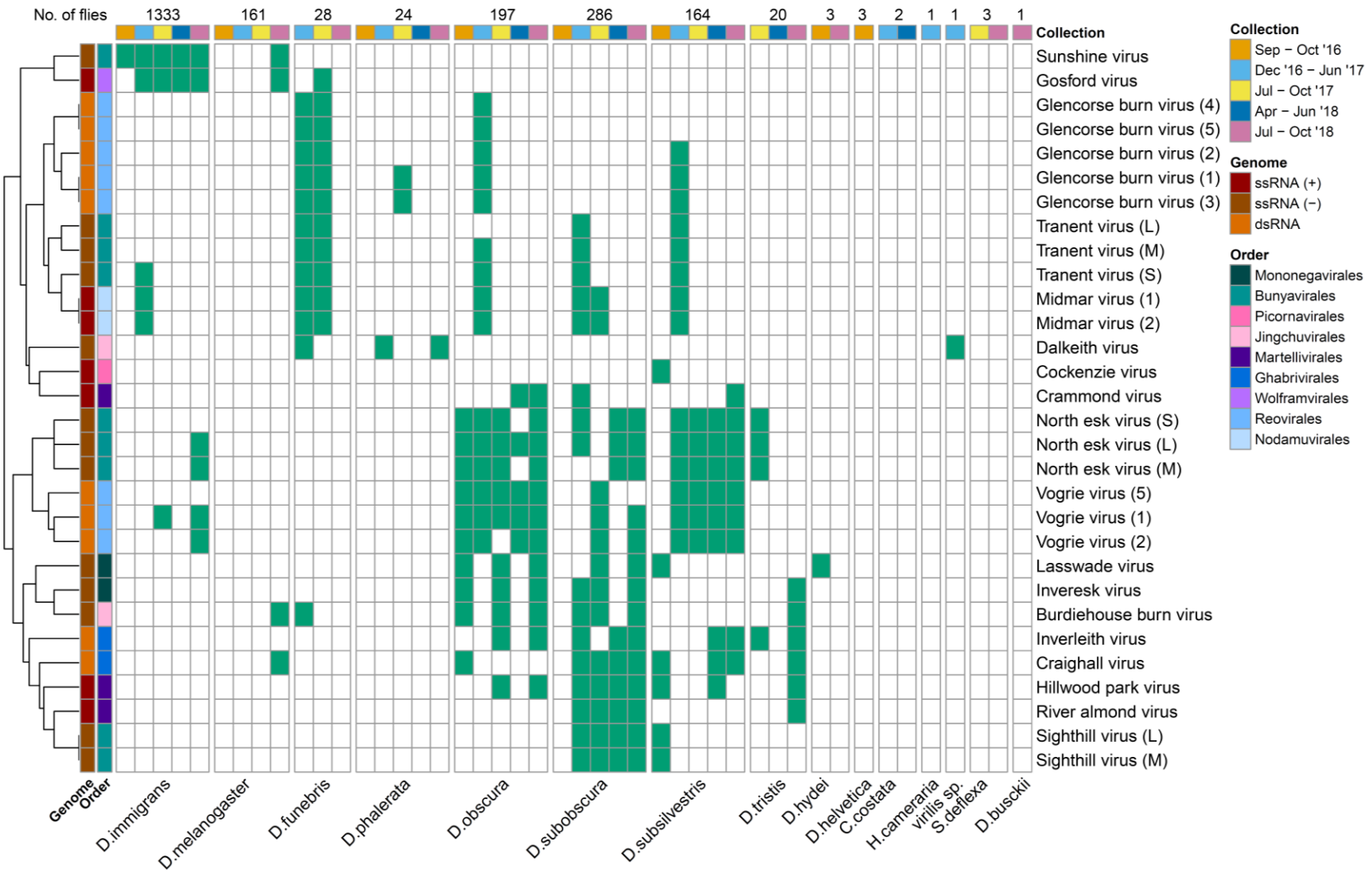


Fig. S2. 3 Host range matrix by RT-PCR for newly described *Drosophila* RNA viruses, including all segments. The plot shows the presence/absence of each of the newly described RNA viruses from this chapter, as measured by RT-PCR on each of the sequenced pools of RNA. The number of each species,

collected across the five pools is denoted along the top of the matrix. The viruses and segments were clustered by the similarity of their presence/absence data, and I used this co-occurrence clustering to support the assignment of segments to the same virus.

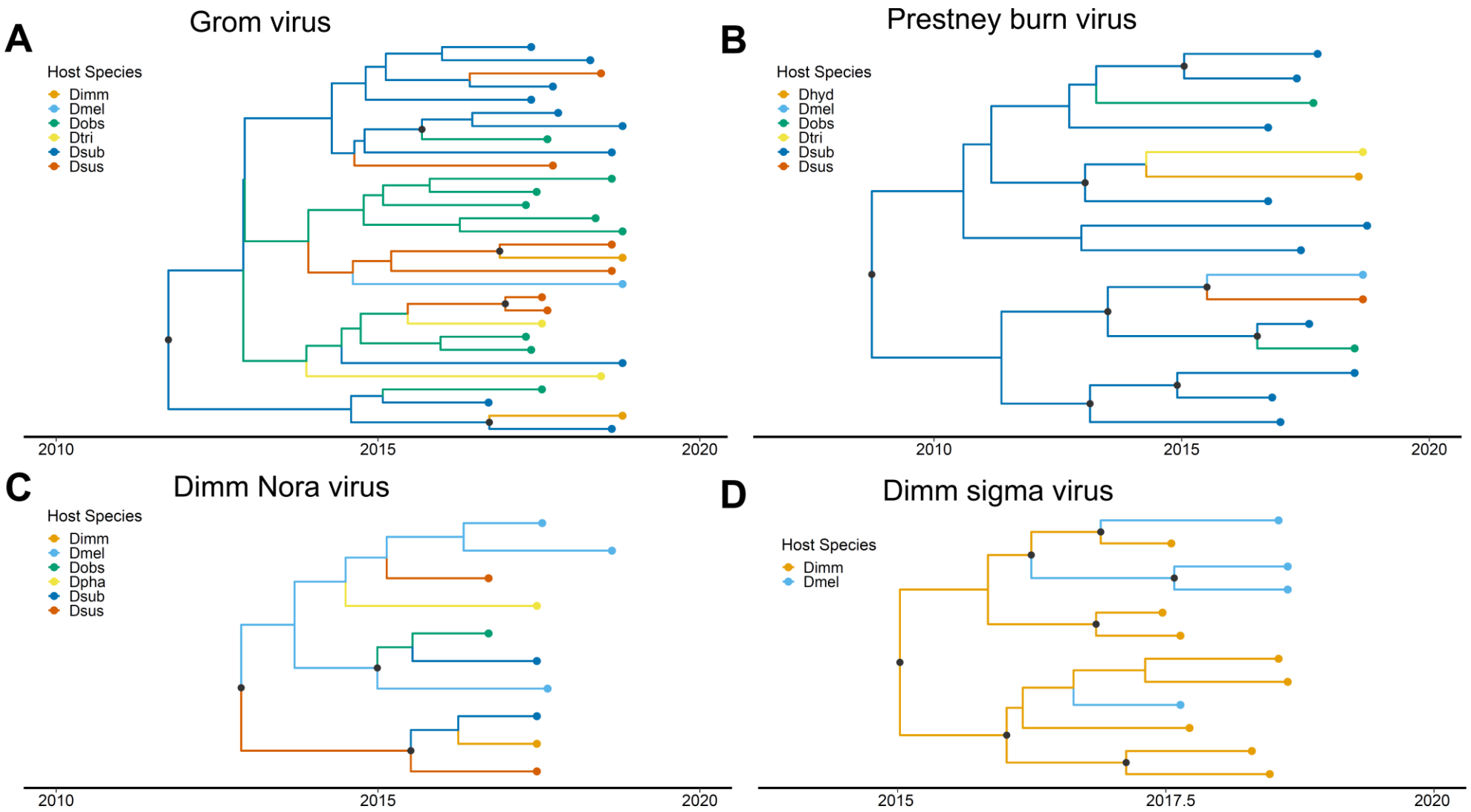


Fig. S2. 4. Phylogenies of multi-host viruses. The plots show the phylogenies of four viruses which infect multiple host species in this system. Branches are coloured by the inferred ancestral host species, and node circles are present if the posterior probability => 0.6.

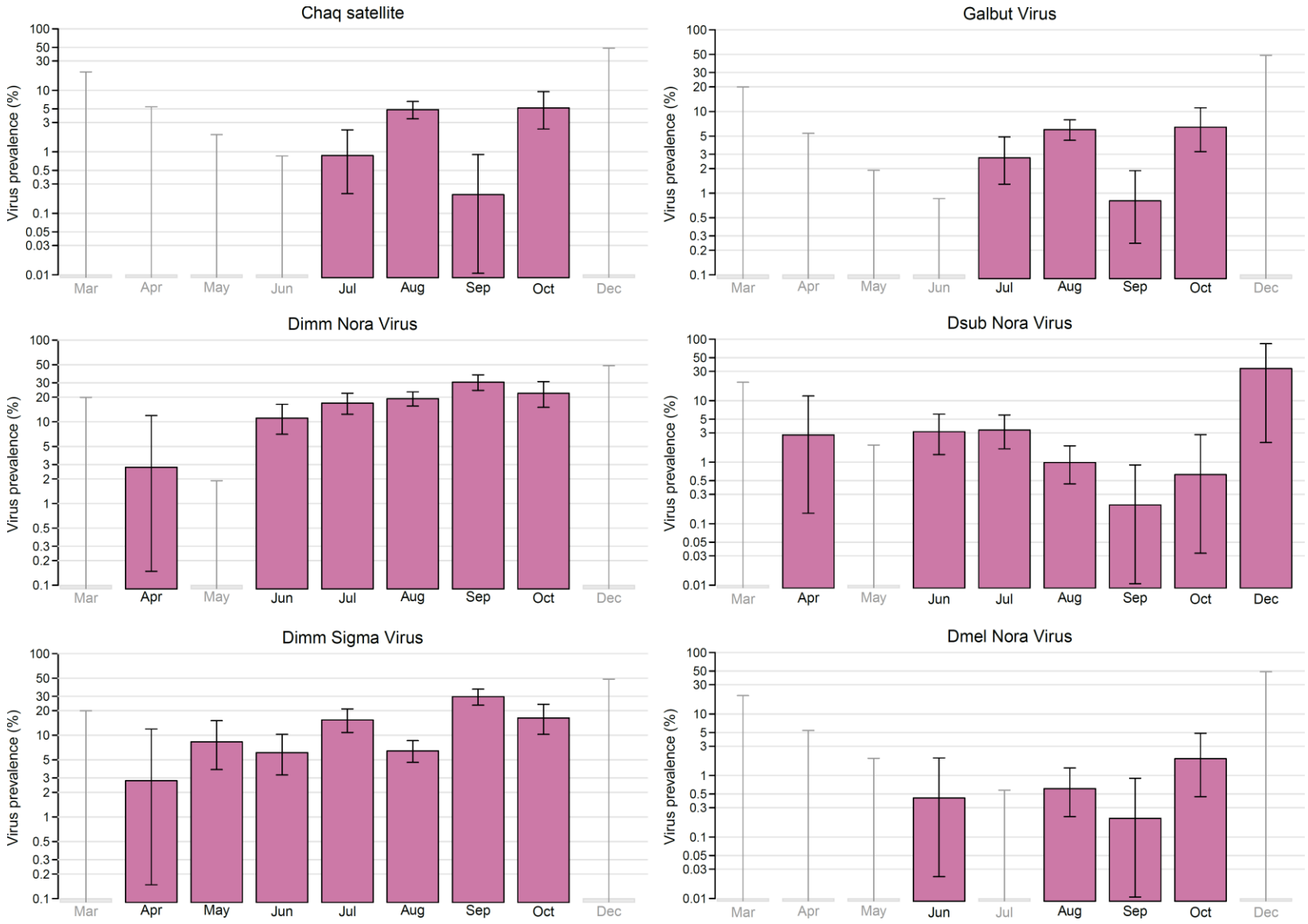
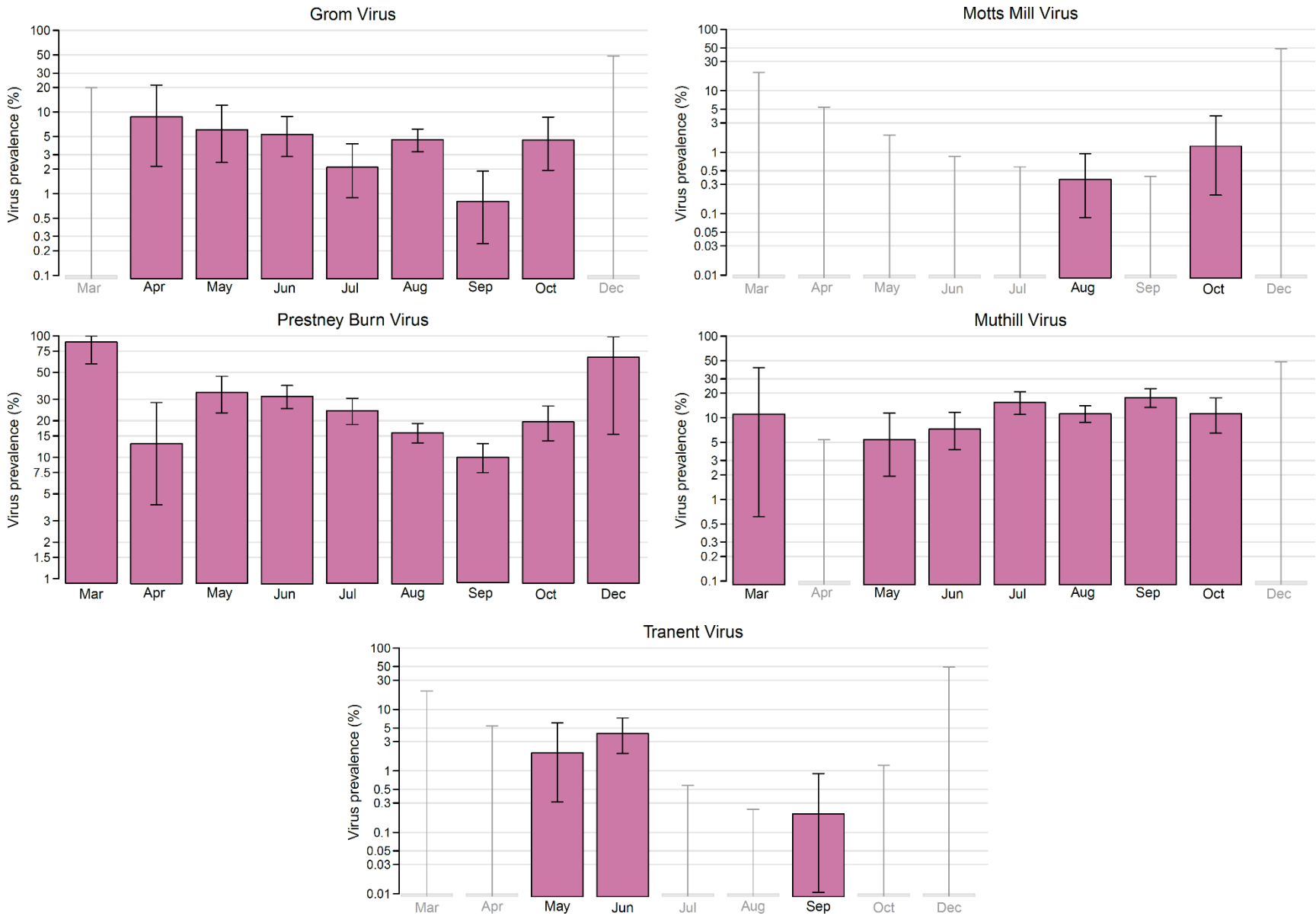


Fig. S2. 5. Prevalence of Drosophila viruses by month (continued on next page). The plots show the estimated individual based prevalence (and 2*log likelihood bounds) for ten Drosophila viruses (and one virus satellite - Chaq), across all years of sampling, and all species, divided by month. Months where no flies were collected are not shown. Greyed out bars indicate that the estimated prevalence was < 0.01 %.


Grom Virus
Virus prevalence (%)
100
50
30
20
10
5
3
2
1
0.5
0.3
0.2
0.1
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Dec
Motts Mill Virus
Virus prevalence (%)
100
50
30
10
5
3
1
0.5
0.3
0.1
0.05
0.03
0.01
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Dec
Prestney Burn Virus
Virus prevalence (%)
100
75
50
30
20
15
10
7.5
5
3
2
1.5
1
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Dec
Muthill Virus
Virus prevalence (%)
100
50
30
20
10
5
3
2
1
0.5
0.3
0.2
0.1
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Dec
Tranent Virus
Virus prevalence (%)
100
50
30
10
5
3
1
0.5
0.3
0.1
0.05
0.03
0.01
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Dec

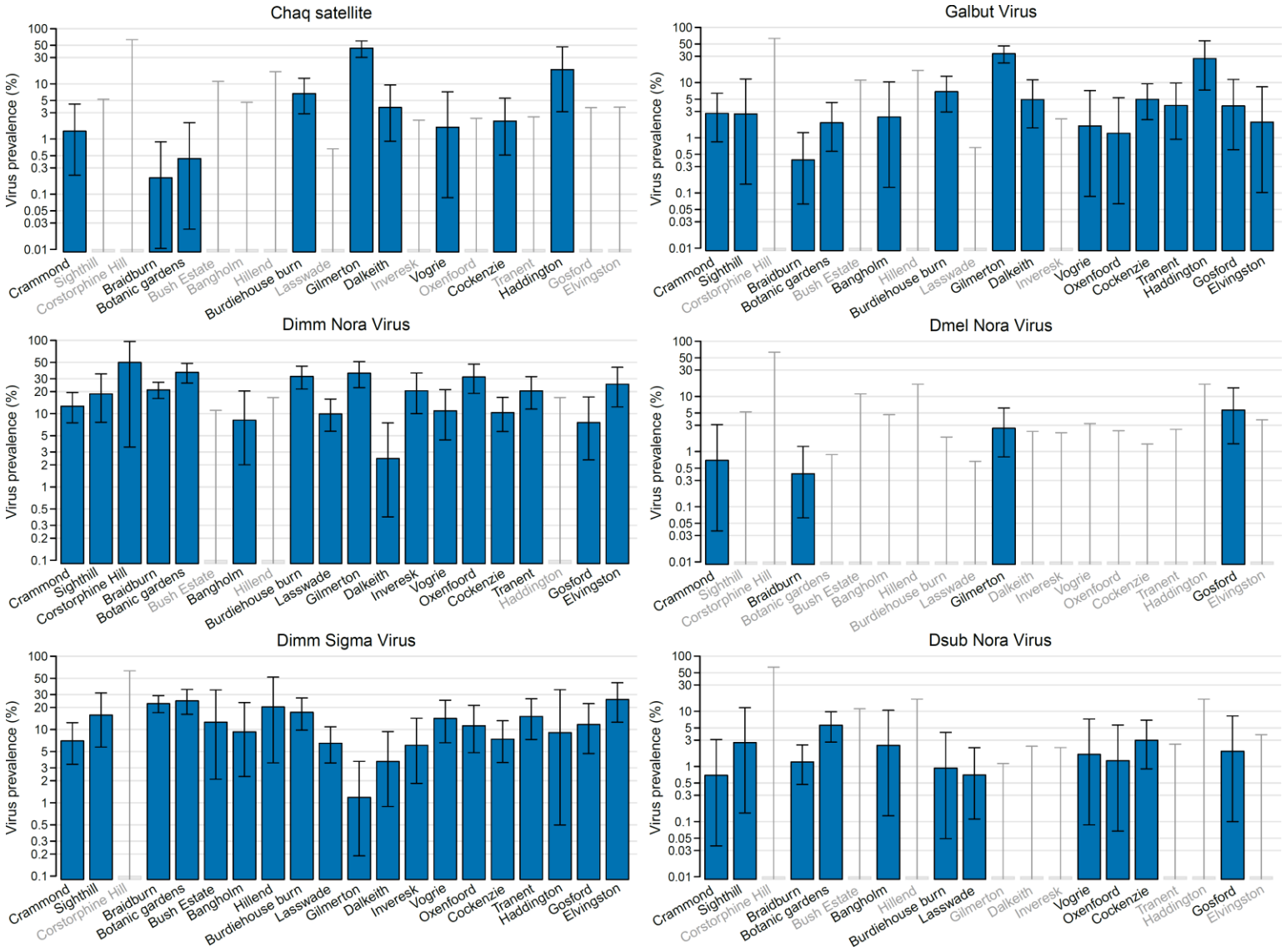
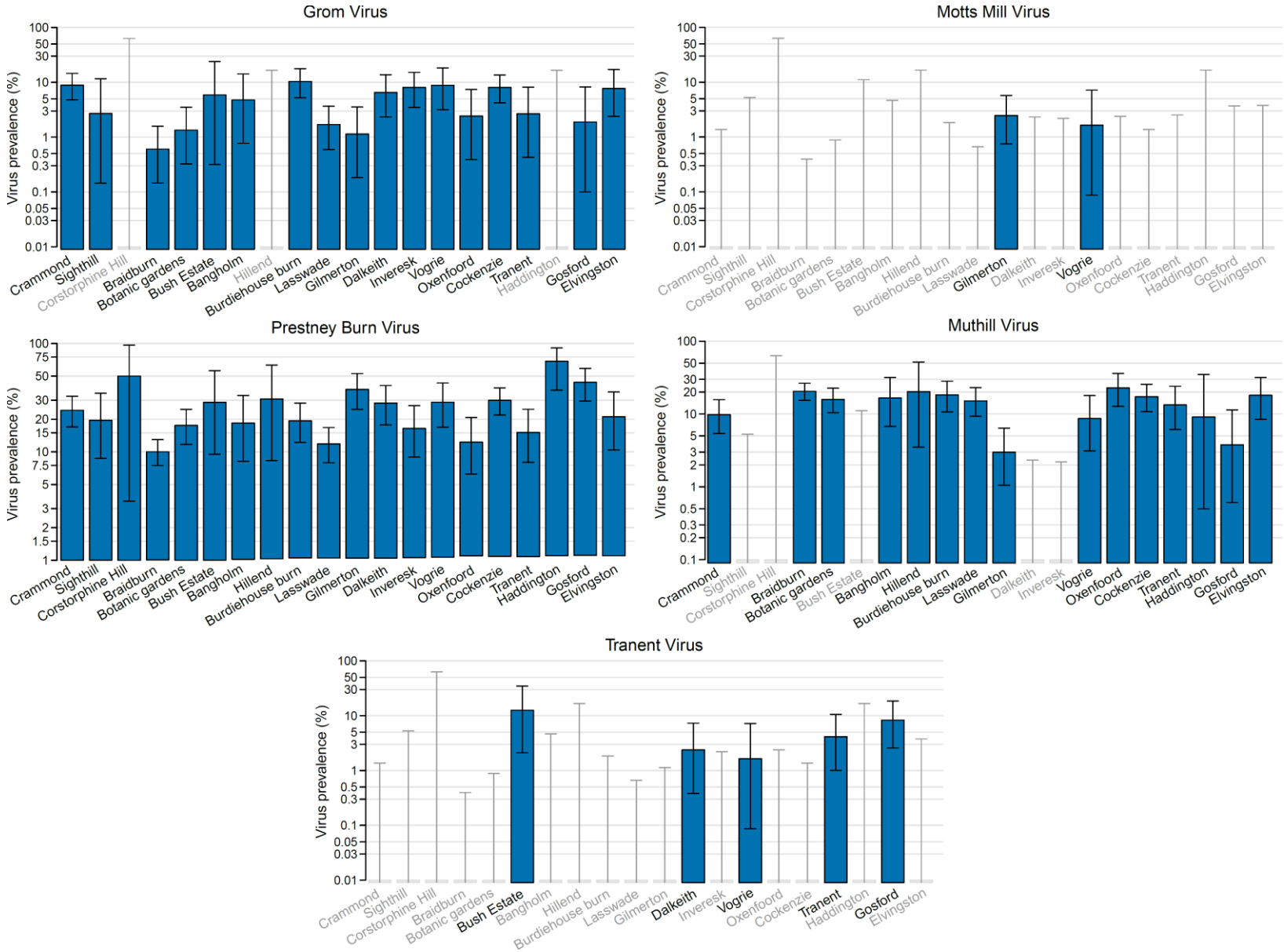


Fig. S2. 6. Prevalence of Drosophila viruses by site (continued on next page). The plots show the estimated individual based prevalence (and 2*log likelihood bounds) for ten Drosophila viruses (and one virus satellite - Chaq), across all years and months of sampling, and all species, divided by site. Sites are ordered by longitude so that the most westerly site is on the far left. Greyed out bars indicate that the estimated prevalence was < 0.01 %.


Grom Virus
Motts Mill Virus
Prestney Burn Virus
Muthill Virus
Tranent Virus
Virus prevalence (%)
Crammond
Sighthill
Corstorphine Hill
Braidburn
Botanic gardens
Bush Estate
Bangholm
Hillend
Burdiehouse burn
Lasswade
Gilmerton
Dalkeith
Inveresk
Vogrie
Oxenfoord
Cockenzie
Tranent
Haddington
Gosford
Elvingston

Contig ID	CDS	Average number of non-synonymous sites	Average number of synonymous sites	Number of non-synonymous SNPs	Number of synonymous SNPs	π_A	π_S	π_A/π_S
KX130344_Kallithea_virus	cds-AQN78547.1	1304.0	454.0	11	2	0.069%	0.056%	1.230
KX130344_Kallithea_virus	cds-AQN78611.1	272.0	100.0	0	1	0.000%	0.078%	0.000
KX130344_Kallithea_virus	cds-AQN78612.1	885.8	296.2	2	1	0.021%	0.014%	1.568
KX130344_Kallithea_virus	cds-AQN78613.1	1091.8	387.2	1	7	0.009%	0.239%	0.039
KX130344_Kallithea_virus	cds-AQN78614.1	396.4	146.6	0	1	0.000%	0.166%	0.000
KX130344_Kallithea_virus	cds-AQN78565.1	875.2	327.8	2	7	0.005%	0.475%	0.011
KX130344_Kallithea_virus	cds-AQN78564.1	528.2	179.8	0	0	0.000%	0.000%	NA
KX130344_Kallithea_virus	cds-AQN78563.1	1074.2	359.8	2	4	0.006%	0.224%	0.025
KX130344_Kallithea_virus	cds-AQN78562.1	813.0	312.0	5	3	0.131%	0.106%	1.237
KX130344_Kallithea_virus	cds-AQN78561.1	2977.0	1004.0	5	7	0.040%	0.123%	0.322
KX130344_Kallithea_virus	cds-AQN78638.1	1410.2	470.8	8	2	0.130%	0.145%	0.891
KX130344_Kallithea_virus	cds-AQN78637.1	321.4	107.6	2	1	0.209%	0.307%	0.679
KX130344_Kallithea_virus	cds-AQN78559.1	222.6	74.4	1	0	0.011%	0.000%	Inf
KX130344_Kallithea_virus	cds-AQN78636.1	1528.2	517.8	3	7	0.026%	0.158%	0.164
KX130344_Kallithea_virus	cds-AQN78551.1	415.0	140.0	5	0	0.259%	0.000%	Inf
KX130344_Kallithea_virus	cds-AQN78635.1	1621.8	574.2	7	1	0.020%	0.009%	2.222
KX130344_Kallithea_virus	cds-AQN78634.1	896.8	306.2	0	3	0.000%	0.150%	0.000
KX130344_Kallithea_virus	cds-AQN78633.1	1260.0	453.0	2	3	0.039%	0.269%	0.145
KX130344_Kallithea_virus	cds-AQN78632.1	305.2	105.8	7	0	0.507%	0.000%	Inf
KX130344_Kallithea_virus	cds-AQN78631.1	563.0	196.0	7	1	0.167%	0.026%	6.342
KX130344_Kallithea_virus	cds-AQN78630.1	828.0	291.0	2	3	0.020%	0.111%	0.182
KX130344_Kallithea_virus	cds-AQN78629.1	2436.4	869.6	15	7	0.031%	0.182%	0.169
KX130344_Kallithea_virus	cds-AQN78628.1	811.2	289.8	2	2	0.022%	0.019%	1.143
KX130344_Kallithea_virus	cds-AQN78627.1	646.6	232.4	1	1	0.019%	0.037%	0.502
KX130344_Kallithea_virus	cds-AQN78626.1	758.0	250.0	1	2	0.005%	0.102%	0.052
KX130344_Kallithea_virus	cds-AQN78625.1	647.6	219.4	0	4	0.000%	0.102%	0.000
KX130344_Kallithea_virus	cds-AQN78624.1	880.6	325.4	4	3	0.096%	0.386%	0.248
KX130344_Kallithea_virus	cds-AQN78622.1	1150.8	394.2	4	3	0.047%	0.019%	2.520
KX130344_Kallithea_virus	cds-AQN78621.1	1175.2	396.8	5	2	0.117%	0.070%	1.677

KX130344_Kallithea_virus	cds-AQN78546.1	606.2	212.8	3	2	0.059%	0.098%	0.604
KX130344_Kallithea_virus	cds-AQN78552.1	699.8	239.2	5	4	0.074%	0.365%	0.201
KX130344_Kallithea_virus	cds-AQN78556.1	707.8	255.2	1	2	0.016%	0.184%	0.086
KX130344_Kallithea_virus	cds-AQN78620.1	2019.6	710.4	5	1	0.041%	0.039%	1.045
KX130344_Kallithea_virus	cds-AQN78619.1	457.8	160.2	0	3	0.000%	0.290%	0.000
KX130344_Kallithea_virus	cds-AQN78618.1	1566.6	545.4	8	9	0.097%	0.344%	0.282
KX130344_Kallithea_virus	cds-AQN78617.1	966.0	333.0	4	5	0.066%	0.269%	0.246
KX130344_Kallithea_virus	cds-AQN78616.1	232.6	79.4	0	1	0.000%	0.627%	0.000
KX130344_Kallithea_virus	cds-AQN78548.1	203.8	72.2	1	0	0.028%	0.000%	Inf
KX130344_Kallithea_virus	cds-AQN78615.1	2193.6	818.4	24	9	0.128%	0.162%	0.794
KX130344_Kallithea_virus	cds-AQN78640.1	810.6	260.4	4	2	0.049%	0.094%	0.528
KX130344_Kallithea_virus	cds-AQN78566.1	433.8	154.2	1	0	0.019%	0.000%	Inf
KX130344_Kallithea_virus	cds-AQN78567.1	554.8	201.2	0	2	0.000%	0.071%	0.000
KX130344_Kallithea_virus	cds-AQN78568.1	1016.8	357.2	3	1	0.050%	0.061%	0.822
KX130344_Kallithea_virus	cds-AQN78569.1	865.0	311.0	3	3	0.040%	0.150%	0.267
KX130344_Kallithea_virus	cds-AQN78570.1	1097.2	396.8	5	4	0.067%	0.213%	0.313
KX130344_Kallithea_virus	cds-AQN78571.1	198.8	68.2	0	1	0.000%	0.199%	0.000
KX130344_Kallithea_virus	cds-AQN78572.1	727.4	250.6	1	1	0.008%	0.016%	0.469
KX130344_Kallithea_virus	cds-AQN78558.1	269.4	93.6	3	0	0.161%	0.000%	Inf
KX130344_Kallithea_virus	cds-AQN78557.1	263.8	87.2	1	0	0.172%	0.000%	Inf
KX130344_Kallithea_virus	cds-AQN78573.1	531.0	201.0	4	1	0.116%	0.107%	1.091
KX130344_Kallithea_virus	cds-AQN78574.1	616.6	208.4	1	0	0.058%	0.000%	Inf
KX130344_Kallithea_virus	cds-AQN78575.1	421.0	158.0	1	0	0.104%	0.000%	Inf
KX130344_Kallithea_virus	cds-AQN78576.1	1272.6	455.4	3	7	0.079%	0.373%	0.212
KX130344_Kallithea_virus	cds-AQN78554.1	1294.6	469.4	10	1	0.198%	0.020%	9.770
KX130344_Kallithea_virus	cds-AQN78623.1	858.0	312.0	8	3	0.201%	0.188%	1.069
KX130344_Kallithea_virus	cds-AQN78577.1	1664.6	561.4	15	7	0.175%	0.267%	0.658
KX130344_Kallithea_virus	cds-AQN78578.1	567.8	206.2	1	5	0.029%	0.485%	0.060
KX130344_Kallithea_virus	cds-AQN78580.1	1282.2	454.8	2	4	0.005%	0.359%	0.015
KX130344_Kallithea_virus	cds-AQN78579.1	641.0	223.0	6	1	0.170%	0.124%	1.375
KX130344_Kallithea_virus	cds-AQN78581.1	748.0	251.0	3	3	0.027%	0.081%	0.327
KX130344_Kallithea_virus	cds-AQN78549.1	488.2	147.8	1	4	0.004%	0.457%	0.009
KX130344_Kallithea_virus	cds-AQN78582.1	3127.0	1064.0	10	6	0.044%	0.095%	0.467

KX130344_Kallithea_virus	cds-AQN78550.1	604.0	218.0	4	4	0.149%	0.165%	0.899
KX130344_Kallithea_virus	cds-AQN78583.1	434.2	156.8	1	0	0.015%	0.000%	Inf
KX130344_Kallithea_virus	cds-AQN78584.1	450.0	153.0	2	0	0.016%	0.000%	Inf
KX130344_Kallithea_virus	cds-AQN78585.1	440.4	150.6	0	1	0.000%	0.096%	0.000
KX130344_Kallithea_virus	cds-AQN78586.1	233.0	94.0	0	2	0.000%	0.283%	0.000
KX130344_Kallithea_virus	cds-AQN78587.1	1068.0	357.0	1	1	0.017%	0.006%	2.951
KX130344_Kallithea_virus	cds-AQN78588.1	483.0	162.0	1	0	0.007%	0.000%	Inf
KX130344_Kallithea_virus	cds-AQN78590.1	1079.6	405.4	4	2	0.012%	0.031%	0.384
KX130344_Kallithea_virus	cds-AQN78589.1	1388.2	492.8	10	4	0.105%	0.123%	0.853
KX130344_Kallithea_virus	cds-AQN78591.1	627.6	224.4	1	1	0.018%	0.173%	0.105
KX130344_Kallithea_virus	cds-AQN78592.1	2341.6	787.4	8	13	0.063%	0.415%	0.153
KX130344_Kallithea_virus	cds-AQN78593.1	546.6	197.4	0	0	0.000%	0.000%	NA
KX130344_Kallithea_virus	cds-AQN78594.1	938.2	342.8	0	6	0.000%	0.238%	0.000
KX130344_Kallithea_virus	cds-AQN78595.1	1284.8	473.2	4	2	0.017%	0.135%	0.130
KX130344_Kallithea_virus	cds-AQN78596.1	794.4	297.6	9	3	0.056%	0.052%	1.085
KX130344_Kallithea_virus	cds-AQN78600.1	2212.8	769.2	2	9	0.008%	0.194%	0.043
KX130344_Kallithea_virus	cds-AQN78599.1	203.4	69.6	0	0	0.000%	0.000%	NA
KX130344_Kallithea_virus	cds-AQN78555.1	565.0	179.0	1	3	0.011%	0.554%	0.019
KX130344_Kallithea_virus	cds-AQN78598.1	214.8	76.2	2	0	0.097%	0.000%	Inf
KX130344_Kallithea_virus	cds-AQN78597.1	1106.0	397.0	2	5	0.029%	0.284%	0.104
KX130344_Kallithea_virus	cds-AQN78601.1	308.4	108.6	0	0	0.000%	0.000%	NA
KX130344_Kallithea_virus	cds-AQN78602.1	782.8	261.2	6	4	0.172%	0.131%	1.313
KX130344_Kallithea_virus	cds-AQN78603.1	624.6	206.4	1	6	0.047%	0.254%	0.185
KX130344_Kallithea_virus	cds-AQN78604.1	877.4	304.6	1	0	0.006%	0.000%	Inf
KX130344_Kallithea_virus	cds-AQN78605.1	117.2	44.8	0	0	0.000%	0.000%	NA
KX130344_Kallithea_virus	cds-AQN78606.1	205.4	79.6	0	0	0.000%	0.000%	NA
KX130344_Kallithea_virus	cds-AQN78607.1	486.6	167.4	2	0	0.076%	0.000%	Inf
KX130344_Kallithea_virus	cds-AQN78608.1	463.4	145.6	1	1	0.032%	0.071%	0.458
KX130344_Kallithea_virus	cds-AQN78609.1	610.6	217.4	4	0	0.078%	0.000%	Inf
KX130344_Kallithea_virus	cds-AQN78639.1	411.2	158.8	3	1	0.219%	0.142%	1.541
KX130344_Kallithea_virus	cds-AQN78560.1	283.2	97.8	2	4	0.184%	0.179%	1.030
KX130344_Kallithea_virus	cds-AQN78610.1	3086.4	1056.6	11	14	0.047%	0.135%	0.350
KX130344_Kallithea_virus	cds-AQN78553.1	1244.4	426.6	7	3	0.139%	0.219%	0.634

KX648536_Linvill_Road_virus	cds-AQN78650.1	1260.2	464.8	12	31	0.163%	1.250%	0.130
KX648536_Linvill_Road_virus	cds-AQN78651.1	1011.5	380.5	11	33	0.130%	1.682%	0.077
Vesanto_virus_UA_Kan_2016_57_S01	hypothetical_protein	378.8	137.3	0	0	0.000%	0.000%	NA
Vesanto_virus_UA_Kan_2016_57_S01	putative_capsid_protein	967.5	310.5	0	0	0.000%	0.000%	NA
Vesanto_virus_UA_Kan_2016_57_S01	DNA_polymerase_B	2461.8	790.3	1	5	0.003%	0.028%	0.110
Vesanto_virus_UA_Kan_2016_57_S02	hypothetical_protein_3	638.3	174.7	0	0	0.000%	0.000%	NA
Vesanto_virus_UA_Kan_2016_57_S02	hypothetical_protein_1	886.7	343.3	0	0	0.000%	0.000%	NA
Vesanto_virus_UA_Kan_2016_57_S02	putative_NS1_protein	1060.0	314.0	2	15	0.005%	0.178%	0.030
Vesanto_virus_UA_Kan_2016_57_S02	putative_coat_protein	481.7	166.3	5	3	0.036%	0.093%	0.390
Vesanto_virus_UA_Kan_2016_57_S02	hypothetical_protein_2	482.3	141.7	1	7	0.013%	0.248%	0.052
Vesanto_virus_UA_Kan_2016_57_S03	putative_DNA_polymerase_B	2606.0	1018.0	52	116	0.431%	2.782%	0.155
Vesanto_virus_UA_Kan_2016_57_S04	putative_structural_protein	2375.0	862.0	2	0	0.002%	0.000%	Inf
Vesanto_virus_UA_Kan_2016_57_S05	putative_nuclease_domain_protein	899.7	273.3	1	1	0.004%	0.012%	0.374
Vesanto_virus_UA_Kan_2016_57_S05	putative_glycoprotein	707.7	219.3	0	0	0.000%	0.000%	NA
Vesanto_virus_UA_Kan_2016_57_S06	hypothetical_protein	1229.0	337.0	14	4	0.036%	0.034%	1.062
Vesanto_virus_UA_Kan_2016_57_S06	putative_glycoprotein	720.0	207.0	4	8	0.017%	0.117%	0.147
Vesanto_virus_UA_Kan_2016_57_S07	hypothetical_protein	1434.0	282.0	26	9	0.040%	0.108%	0.366
Vesanto_virus_UA_Kan_2016_57_S08	hypothetical_protein_1	561.3	161.7	0	0	0.000%	0.000%	NA
Vesanto_virus_UA_Kan_2016_57_S09	putative_NACHT_domain_protein	1003.8	370.2	1	1	0.019%	0.045%	0.421
Vesanto_virus_UA_Kan_2016_57_S10	putative_glycoprotein	587.8	195.3	0	0	0.000%	0.000%	NA
Vesanto_virus_UA_Kan_2016_57_S10	hypothetical_protein	1179.5	404.5	5	1	0.076%	0.124%	0.613
Vesanto_virus_UA_Dro_2016_56_S01	hypothetical_protein	487.5	175.5	0	0	0.000%	0.000%	NA
Vesanto_virus_UA_Dro_2016_56_S01	putative_capsid_protein	966.5	311.5	0	41	0.000%	1.283%	0.000
Vesanto_virus_UA_Dro_2016_56_S01	DNA_polymerase_B	2459.8	792.3	38	170	0.146%	2.077%	0.070
Vesanto_virus_UA_Dro_2016_56_S02	hypothetical_protein_3	614.0	199.0	0	0	0.000%	0.000%	NA
Vesanto_virus_UA_Dro_2016_56_S02	hypothetical_protein_1	861.3	368.8	0	0	0.000%	0.000%	NA
Vesanto_virus_UA_Dro_2016_56_S02	putative_NS1_protein	1026.0	360.0	1	6	0.008%	0.338%	0.023
Vesanto_virus_UA_Dro_2016_56_S02	putative_coat_protein	466.0	182.0	6	11	0.098%	0.654%	0.150
Vesanto_virus_UA_Dro_2016_56_S02	hypothetical_protein_2	463.5	160.5	3	11	0.032%	0.519%	0.062
Vesanto_virus_UA_Dro_2016_56_S03	putative_DNA_polymerase_B	2691.0	933.0	108	212	1.018%	4.840%	0.210
Vesanto_virus_UA_Dro_2016_56_S04	putative_structural_protein	2373.6	863.4	12	19	0.034%	0.178%	0.190
Vesanto_virus_UA_Dro_2016_56_S05	putative_nuclease_domain_protein	818.6	354.4	3	7	0.014%	0.197%	0.072
Vesanto_virus_UA_Dro_2016_56_S05	putative_glycoprotein	642.3	284.8	2	4	0.078%	0.068%	1.147

Vesanto_virus_UA_Dro_2016_56_S06	hypothetical_protein	1157.2	408.8	6	11	0.076%	0.417%	0.182
Vesanto_virus_UA_Dro_2016_56_S06	putative_glycoprotein	672.4	254.6	4	3	0.113%	0.407%	0.276
Vesanto_virus_UA_Dro_2016_56_S07	hypothetical_protein	1337.3	378.7	4	5	0.042%	0.485%	0.086
Vesanto_virus_UA_Dro_2016_56_S08	hypothetical_protein_1	542.3	186.8	32	48	1.131%	3.938%	0.287
Vesanto_virus_UA_Dro_2016_56_S09	putative_NACHT_domain_protein	1103.7	270.3	4	1	0.011%	0.144%	0.080
Vesanto_virus_UA_Dro_2016_56_S10	putative_glycoprotein	548.4	234.6	4	5	0.107%	0.484%	0.220
Vesanto_virus_UA_Dro_2016_56_S10	hypothetical_protein	1144.5	484.5	6	9	0.103%	0.251%	0.412
Vesanto_virus_UA_Dro_2016_56_S11	putative_DNA_polymerase_B	1268.2	429.8	25	37	0.835%	3.927%	0.212
Vesanto_virus_UA_Dro_2016_56_S11	putative_DNA_polymerase_B_2	1555.4	562.6	14	63	0.434%	4.124%	0.105
Vesanto_virus_FR_Got_2015_48_S02	hypothetical_protein_3	612.3	200.8	1	1	0.008%	0.175%	0.048
Vesanto_virus_FR_Got_2015_48_S02	hypothetical_protein_1	859.3	370.8	0	0	0.000%	0.000%	NA
Vesanto_virus_FR_Got_2015_48_S02	putative_NS1_protein	1022.3	360.8	7	9	0.052%	0.486%	0.106
Vesanto_virus_FR_Got_2015_48_S02	putative_coat_protein	465.8	182.3	7	6	0.295%	1.173%	0.251
Vesanto_virus_FR_Got_2015_48_S02	hypothetical_protein_2	260.8	87.3	9	3	0.703%	0.440%	1.598
Vesanto_virus_FR_Got_2015_48_S03	putative_DNA_polymerase_B	2692.6	931.4	92	188	1.149%	7.328%	0.157
Vesanto_virus_FR_Got_2015_48_S04	putative_structural_protein	2433.5	803.5	8	11	0.023%	0.042%	0.548
Vesanto_virus_FR_Got_2015_48_S05	putative_nuclease_domain_protein	901.3	271.7	1	0	0.016%	0.000%	Inf
Vesanto_virus_FR_Got_2015_48_S05	putative_glycoprotein	707.7	219.3	0	1	0.000%	0.180%	0.000
Vesanto_virus_FR_Got_2015_48_S06	hypothetical_protein	1185.0	381.0	0	0	0.000%	0.000%	NA
Vesanto_virus_FR_Got_2015_48_S06	putative_glycoprotein	619.5	214.5	1	1	0.071%	0.040%	1.780
Vesanto_virus_FR_Got_2015_48_S09	putative_NACHT_domain_protein	1079.3	327.8	6	10	0.151%	0.672%	0.224
Vesanto_virus_FR_Got_2015_48_S10	putative_glycoprotein	540.5	242.5	4	5	0.264%	0.802%	0.329
Vesanto_virus_FR_Got_2015_48_S10	hypothetical_protein	1131.0	498.0	7	6	0.188%	0.381%	0.494
Vesanto_virus_FR_Got_2015_48_S12	putative_NACHT_domain_protein	1048.3	325.8	42	39	0.873%	2.103%	0.415
Vesanto_virus_FR_Got_2015_48_S12	hypothetical_deoxynucleoside_kinase_protein	543.3	161.8	11	12	0.407%	1.361%	0.299
Vesanto_virus_2015_01_S10	putative_glycoprotein	588.3	194.8	17	23	0.285%	1.383%	0.206
Vesanto_virus_2015_01_S10	hypothetical_protein	1209.8	410.3	32	36	0.220%	0.814%	0.270
Vesanto_virus_2014_38_S07	hypothetical_protein	1363.3	382.7	11	5	0.136%	0.290%	0.471
Vesanto_virus_2015_10_S03	putative_DNA_polymerase_B	2640.3	983.7	109	225	0.756%	5.229%	0.145

Table S3. 1 Total population πA, πS and πA/πS for each gene in the genomes of Kallithea, Linvill Road and Vesanto virus (including all haplotypes). Also includes the number of synonymous and non-synonymous SNPs identified in each gene by popoolation (Kofler et al. 2011a) filtered to only include those with a minor allele frequency of 1% or above.

Virus	Virus Classification	First Description	Genbank Accession No.	Forward primer name	Forward primer sequence (5'-3')	Reverse primer name	Reverse primer sequence (5'-3')	Product length (bp)	Tm (°C)	Extension time	Primer source
Chaq virus	Unclassified	(Webster et *al*., 2015)	KP714088	222_140F	AACAGAACGWCTGCTTTTTGGAAATCC	222_660R	TCCATGTCCTGTHGGGTCTATCTG	520	60	45	(Webster et *al*., 2015)
Cherry Gardens virus	*Rhabdoviridae* (-ssRNA)	(Webster et *al*., 2016)	KU754524	Dsub_CherryGardens_03600F	CTGGTGATGGAACTGGTGGAG	Dsub_CherryGardens_04416R	TGGTCGTGGAGCCTGATTAG	814	59	60	Nathan Medd (Obbard lab) thesis
Craigie's Hill virus	Alphanodavirus (+ssRNA Segmented)	(Webster et *al*., 2015)	KP714084 KP714085	NV-L_1.69F	GCAAAATCCGTGGTTCATACCAG	NV-L_2.89R	CTTAACAGGACGCTCCAAGTGGAT	1200	60	90	(Webster et *al*., 2015)
				674_100F	CCTATCTGTCAAGCTGTWCTGCCAAC	674_1540R	GTGTGCCAACTAGGCTCAGGAG	1440	62	90	
Dimm Nora virus	Picornavirales (+ssRNA)	(van Mierlo et *al*., 2014)	KF242511	Dimm_Nora_A_10358F	TGAATCCTTGGATGCGAACGG	Dimm_Nora_A_11173R	TTGATGCTGCTCCTAACACGG	816	61	60	Nathan Medd (Obbard lab) thesis
				Dimm_Nora_B_03392F	AGGCATTACACGCTCGATTCC	Dimm_Nora_B_04071R	TTCACTCGCCATAGGAACACC	680	61	60	
Dimm Sigma virus (L & N genes)	Sigmavirus (Rhabdoviridae) (-ssRNA)	(Longdon et al., 2011)	KR822814	DImmSV_L_F2	TGATAGATCCTTCCGCCATC	DImmSV_L_R2	GAATGCTCCAACCCTTTTGA	696	56	60	Longdon et al. 2010
				DImmSV_N_ F1	CTAGCATTTGCCGGGATAAA	DImmSV_N_R1	AATGCATTCCTGGTCTTTGG	782	56	60	
Galbut virus	Unclassified partitivirus	(Webster et *al*., 2015)	KP714100 KP714099	407_170F	GATCGAGATGGAACTCCRCTCTC	407_750R	GCCKCATACTTGGTGCTGCCAACTG	580	60	60	(Webster et *al*., 2015)
Muthill virus	cf. Negevirus (+ssRNA)	(Webster et *al*., 2016)	KU754517	Dsuz_Muthill_06921F	AATGCGTTTCCTAGCCCACAC	Dsuz_Muthill_07613R	ATGGCGCGTTGTTTGTTTCG	693	61	60	Nathan Medd (Obbard lab) thesis
Pow burn virus	Picornaviridae (+ssRNA)	(Webster et *al*., 2016)	KU754519	Dobs_PowBurn_B_03155F	CACAAGAAGAAGCGTGACTC	Dobs_PowBurn_B_03926R	TGTTAGCCTCTCCGTATGC	772	55.5	60	Nathan Medd (Obbard lab) thesis
Corseley	Unclassified (+ssRNA)	(Webster et *al*., 2016)	KU754520	DsubDsuz_Corseley_00511F	ACGTGTTGAGCGAGGAGTAC	DsubDsuz_Corseley_01311R	TTCTGCTACACTCATGCTGGC	801	60	60	Nathan Medd (Obbard lab) thesis
Craigmillar park virus	Alphanodavirus (+ssRNA Segmented)	(Webster et *al*., 2016)	KU754525 KU754526	Dsus_Craigmillar_seg1_01679F	AGCAGTATCCGTGGTTCATCC	Dsus_Craigmillar_seg1_02483R	TCATGTTGAGGCCAGGAATCAG	805	59	60	Nathan Medd (Obbard lab) thesis
				Dsus_Craigmillar_seg2_00112F	GTGCCGACTAAGGTTGCTCTC	Dsus_Craigmillar_seg2_00806R	CGTGTGAGTTGATTCCACAGAC	695	59	60	Nathan Medd (Obbard lab) thesis
Grom virus	cf. Sobemoviruses and	(Webster et *al*., 2016)	KU754506	Dobs_Grom_00728F	CGGCAGGTCACAACATCTCAT	Dobs_Grom_01554R	GCTTTGGGTGACTGTGGACT	826	60	60	Nathan Medd (Obbard lab) thesis

	Poleroviruses (+ssRNA)										
La Jolla virus	Iflavirus (+ssRNA)	(Webster et *al*., 2015)	KP714073	SB-L_6.77_F	GTGGAGTAAAGCAACGACTTGG	SB_ASSAY_1R	CAACTGCRTGTTTGAGTTCCCAACGA	1300	60	90	(Webster et al., 2015)
Lye green virus	Rhabdoviridae (-ssRNA)	(Webster et *al*., 2016)	KU754522	Dobs_Lye_Green_04753F	TTGTCGAGAATAGCAGGAGTCC	Dobs_Lye_Green_05516R	AGTCCGGTCCTAGTCTGAAGC	764	59	60	Nathan Medd (Obbard lab) thesis
Motts mill virus	cf. Sobemoviruses and Poleroviruses (+ssRNA)	(Webster et *al*., 2015)	KP714076 KP714077	Luteo_1F_364	AATAAATCATAAATCGTGCTTGTTTCCTTGGC	Luteo_2R_1758	AATAAATCATAAGGTTGAACCAGTCGGTGAAT	1400	67	60	(Webster et al., 2015)
Dmel Nora virus	Picornavirales (+ssRNA)	(Habayeb et *al*., 2006)	NC_007919	Nora_6220F	GACCATTGGCACAAATCACCATTTG	Nora_7210R	TCTTAGGCCGGTTGTCTTCACCC	990	60	60	(Webster et al., 2015)
Prestney burn virus	cf. Sobemoviruses and Poleroviruses (+ssRNA)	(Webster et *al*., 2016)	KU754507	Dsub_Prestney_Burn_A_00190F	GGCCAATTGACTGAATCGGACC	Dsub_Prestney_Burn_A_00880R	TGGTGGTGGTTGGTTGATCG	690	61	60	Nathan Medd (Obbard lab) thesis
***Drosophila* A virus**	cf. Permutotetraviridae (+ssRNA)	(Plus et *al.,* 1975)	NC_012958	DAV_3300F	TGCAAGTAAGCTCTTGCCAACCCT	DAV_3940R	AGATACCACTTACGGGTGGTTGC	640	60	60	(Webster et al., 2015)
D.buskii Rhabdovirus	Rhabdoviridae (-ssRNA)	(Longdon et al., 2015)	KR822813 KR822814	Dbus_rhab_13126F	GAATTGTGCATCGGTGGAGG	Dbus_rhab_13617R	TTGATGCCACAGGAGTTAGGC	491	59	60	Nathan Medd (Obbard lab) thesis
Kinkell virus	Iflavirus (+ssRNA)	(Webster et *al*., 2016)	KU754510	Dsus_Kinkell_03036F	TGTTGTGTACGAGCTGTGGTC	Dsus_Kinkell_03777R	ACACCGATGAGAGCGAGGATG	741	59	60	Nathan Medd (Obbard lab) thesis
Eccles virus	Reoviridae (dsRNA Segmented)	(Medd et al., 2018)	MF893265: MF893270	Eccles_virus_seg1_00828F	CCGAAGAGTTGCGAGTTGG	Eccles_virus_seg1_01538R	CGTCAATGAGTTCCTGAGGC	711	58.5	60	Nathan Medd (Obbard lab) thesis
				Eccles_virus_seg4_02305F	TACTATGCTACTCGTCGCGC	Eccles_virus_seg4_03050R	CAATTGTTAGCTCCACCCGG	745	59	60	Nathan Medd (Obbard lab) thesis
Buckhurst virus	cf. Negevirus (+ssRNA)	(Webster et *al*., 2016)	KU754516	DobsDsub_Buckhurst_06402F	TTCGGAGGTGGTGATAACGC	DobsDsub_Buckhurst_07190R	AGCAGCTTGTGTATCCAACC	738	58	60	Nathan Medd (Obbard lab) thesis
Withyham virus	Rhabdoviridae (-ssRNA)	(Webster et *al*., 2016)	KU754523	Dobs_Withyham_00425F	CTGGGCACTTTGGAATCCTC	Dobs_Withyham_00970 R	ATGTGTCCGCCATCATCAAC	545	57.5	60	Nathan Medd (Obbard lab) thesis
Larkfield virus	Totiviridae (dsRNA)	(Medd et *al*., 2018)	MF893249	DrosMix_Totivirus_01124F	GTACCAGATCATTGCTATGACC	DrosMix_Totivirus_02065R	GGAATCGTGTATATCGAAGAGC	941	61	60	Medd et al., 2018
Thika virus	Picornavirales (+ssRNA)	(Webster et *al*., 2015)	KP714072	Cripa_C3_5625F	CTTCGAAGCATCYCTGCATCGTAAAG	Cripa_univ_6560R	GCACCCACAGCTAGCATRTCTGG	900	60	60	(Webster et al., 2015)
Bloomfield virus	Cypovirus (Reoviridae)	(Webster et *al*., 2015)	KP714090-KP714098	Reo1_S4_1F_620	AATAAATCATAACTATGGTTATCGATTGCATGGTCC	Reo1_S4_2R_1636	AATAAATCATAAGTAAACAAATCAAAACCATC	1000	58	60	(Webster et *al*., 2015)

	(dsRNA Segemented)			Reo1_S7_qPCR_2F_494	AATAAATCATAAATTTTTGGACTCAGATTGG	Reo1_S7_qPCR_4R_1397	AATAAATCATAAGCCAAAATACTTGTTCCAG	900	65	60	
***Drosophila* C virus**	Cripavirus (Dicistroviridae) (+ssRNA)	(Jousset et *al*., 1972)	NC_001834	DCV7	AGTATGATTTTGATGCAGTTGAATCTC	DCV8	GAAGCACGATACTTCTTCCAAACC	524	59.5	60	(Kapun et al., 2010)
***Drosophila melanogaster* American Nodavirus**	Nodavirus (+ssRNA)	(Wu et *al*., 2009)	GQ342966 GQ342965	Dmel_ANV_2_00373F	CACCTGGAGTCGTTATCTGG	Dmel_ANV_2_01021R	GGCCGAATTGATACAGCATAGC	695	58	60	Nathan Medd (Obbard lab) thesis?
Drosophila X virus	Entomobirnavirus (dsRNA)	(Teninges et *al*., 1979)	NC_004177 NC_004169	DXV_SegB_F	CATGCCAATCAGTGATGACG	DXV_SegB_R	GCTGTTGTCATTGCCACATC	738	57.7	60	Nathan Medd (Obbard lab) thesis
Charvil virus	Flaviviridae	(Webster et *al*., 2015)	KP714089	Dmel_Charvil_01343F	AGGAGTTGACGGATGATGAGG	Dmel_Charvil_01868R	GTTGGTGCGTACTTGTGAACC	525	59	60	Nathan Medd (Obbard lab) thesis?
Kallithea virus	Nudivirus (dsDNA)	(Webster et *al*., 2015)	KP714101-KP714108	NudiPif1_F	CGACATCACATTCGACCCATATCC	NudiPif1_R	TCCCATAAAGTGCGATCCCATAG	970	62	90	(Webster et *al*., 2015)
Dansoman virus	Unclassified (+ssRNA)	(Webster et *al*., 2015)	KP714086 KP714087	CBP1-L_0.90F	GCGCAGACGGAGGACGGCA	CBP1-L_2.27R	ARCGGKGTCACWCGCGGCTC	1300	66	60	(Webster et *al*., 2015)
Kilifi virus	Picornavirales (+ssRNA)	(Webster et *al*., 2015)	KP714071	Dmel_Kilifi_A_03173F	CTGGTCGCTGTCATGAGGTTG	Dmel_Kilifi_A_03875R	ATTGGACGCTGTGATGTCGC	703	60	60	Nathan Medd (Obbard lab) thesis?
Machanay virus	Picornaviridae (+ssRNA)	(Webster et *al*., 2016)	KU754504	Dobs_Machany_02667 F	ATGTTGGCAGTGGGAGCAATC	Dobs_Machany_03666R	TTGTGGTTGAGGTTGATGGGC	1000	60	90	Nathan Medd (Obbard lab) thesis?
Empeyrat virus	Cripavirus (+ssRNA)	(Webster et *al*., 2016)	KU754505	Sdef_Empeyrat_04276F	ACAACGACAATGGACAGGTG	Sdef_Empeyrat_05005R	GTCGATCTACATACTGCGGAC	730	57	60	Nathan Medd (Obbard lab) thesis
Newington virus	Alphanodavirus (+ssRNA Segmented)	(Webster et *al*., 2016)	KU754529 KU754530	Dimm_Newington_RNA1_00464F	TTCCGCAACGATCAACCAGC	Dimm_Newington_RNA1_01221R	TGGACGCGTGGCATTGTAG	758	60	60	Nathan Medd (Obbard lab) thesis?
Eridge virus	Entomobirnavirus (dsRNA Segmented)	(Webster et *al*., 2016)	KU754527 KU754528	Dimm_Eridge_segA_00638F	CTCTGCAATCACCAACGCAC	Dimm_Eridge_segA_01403R	TGCCAGAAGTCGAAGCTTGC	766	60	60	Nathan Medd (Obbard lab) thesis
				Dimm_Eridge_segB_02169F	GAGGTAATGAGCCTGCTAAGCC	Dimm_Eridge_segB_02997R	CTGTTGATGACTTGCGTGCG	828	60	60	
Torrey pines virus	Reoviridae (dsRNA Segemented)	(Webster et *al*., 2015)	KP714078-KP714083	345_425F	GACGTCVTACATCAACGCTAACACGG	345_905R	CACGACTGCAGGAGCATCATTAAC	480	66	60	(Webster et *al*., 2015)
Soudat virus	Cypovirus (dsRNA segmented)	(Webster et *al*., 2016)	KU754531–KU754534	Sdef_Soudat_seg3_02483F	TCAACTAGCGCAATTATGGAAGC	Sdef_Soudat_seg3_03300R	ACCATCCATCCACAATTCATCACC	818	60	60	Nathan Medd (Obbard lab) thesis

Grange virus	Reoviridae (dsRNA Segemented)	(Webster et *al.*, 2016)	KU754536–KU754538	Dsub_Grange_seg1_00409F	TGACACGGCGATATCCAGATC	Dsub_Grange_seg1_01126R	ACTCCAACTCTTGTGCTCAACC	718	60	60	Nathan Medd (Obbard lab) thesis
				Dsub_Grange_seg2_00962 F	CGTGTGATGGCTTGGTTGTC	Dsub_Grange_seg2_01724R	CGCTTGTCTGTATAGGTCTCGG	768	59	60	Nathan Medd (Obbard lab) thesis
Hermitage virus	Unclassified (RNA)	(Webster et *al.*, 2016)	KU754511 KU754512	Dimm_Hermitage_A_00082F	ACATGTATCAACCACCGCGAC	Dimm_Hermitage_A_00845R	ACCGATTTGACACCAGGCTTG	763	60	60	Nathan Medd (Obbard lab) thesis
				Dimm_Hermitage_B_01531F	TTGAAGGTGACGGAAGCCATC	Dimm_Hermitage_B_02182R	CGTTCTTGGTGTGCTCATCG	651	59	60	Nathan Medd (Obbard lab) thesis
Takaugu virus	Unclassified (RNA)	(Webster et *al.*, 2016)	KU754513 KP757925	Dmel_Takaungu_00559F	GGTGACCTAGAAGTTCCGCATG	Dmel_Takaungu_01375R	TGAGCATCACCAGTCCTGC	816	59	90	Nathan Medd (Obbard lab) thesis
Braid Burn virus	cf. Poleroviruses and Sobemoviruses (+ssRNA)	(Webster et *al.*, 2016)	KU754508	Dsus_BraidBurn_A_00264F	TTGTTATGATCGGCTGCGCA	Dsus_BraidBurn_A_00646R	AATGAAAGGCCCGTTGGTGT	383	60	45	Nathan Medd (Obbard lab) thesis
La Tardoire virus	cf. Poleroviruses and Sobemoviruses (+ssRNA)	(Webster et *al.*, 2016)	KU754509	Sdef_La_Tardoire_00223F	ATGACGGATGGTTGGTTGGAC	Sdef_La_Tardoire_00918R	AGACCGGATTAACGCTCTCC	696	58	60	Nathan Medd (Obbard lab) thesis
Twyford virus	Iflaviriidae (+ssRNA)	(Webster et *al.*, 2015)	KP714075	SB-L_1.51_F	CGCAGTCAGTTTGCATCAGG	SB.TWY_2.57_R	CTCAGCTAAGGAGCCTTCCAT	900	62	90	(Webster et al., 2015)
Tartou virus	Unclassified (+ssRNA)	(Webster et *al.*, 2016)	KU754521	Sdef_Tartou_00404F	CGCATTGAATACGCCAGAACC	Sdef_Tartou_01080R	GCTATCCGAGACATGTGTTGC	677	59	60	Nathan Medd (Obbard lab) thesis
Blackford virus	cf. Negevirus (+ssRNA)	(Webster et *al.*, 2016)	KU754514	Dtri_Blackford_01878F	ATCAAGCCGTCCGTGGAATC	Dtri_Blackford_02525R	TCCGCACACAATGTCCTTCG	648	59	60	Nathan Medd (Obbard lab) thesis
Berkley virus	Picornavirales (+ssRNA)	(Webster et *al.*, 2015)	SRA SRR070416	Dmel_Berkeley_02000F	GGTAAGGCTGGATGCTTGGT	Dmel_Berkeley_03057R	AGGGAGACGCAAGCATTTGA	1058	60	90	Nathan Medd (Obbard lab) thesis
Bofa virus	cf. Negevirus (+ssRNA)	(Webster et *al.*, 2016)	KU754515	Dmel_Bofa_05881F	CGGAAGCAGCAACATTGAGAG	Dmel_Bofa_06634R	GGTTCCAATAGTCGCGGTC	754	58	60	Nathan Medd (Obbard lab) thesis
Brandeis virus	Unclassified cf. Negevirus and Virgaviridae (+ssRNA)	(Webster et *al.*, 2015)	SRA SRR486227	Dsuz_Brandeis_05938F	TCTTCAACCGCATGTCCGTG	Dsuz_Brandeis_06869R	GGTGAAGGTGGTGGCATGAC	932	60	90	Nathan Medd (Obbard lab) thesis
Marsac virus	cf. Negevirus (+ssRNA)	(Webster et *al.*, 2016)	KU754518	Sdef_Marsac_04640F	ACATGTGTCCGCCAAGCTAC	Sdef_Marsac-_05733R	TCGTTCAAGTCGCGTGTGAG	1093	60	90	Nathan Medd (Obbard lab) thesis
Versanto virus	Bidensovirus (ssDNA)	(Kapun et *al.*, 2018)	KX648533 KX648534	Dmel_Vesanto_02080F	ATTGCAGACGACGACACCAC	Dmel_Vesanto_02920R	CCTTGACACGCTTACCATGC	841	60	60	Nathan Medd (Obbard lab) thesis

Viltain virus	Densovirus (ssDNA)	(Kapun et *al*., 2018)	KX648535	Dmel_Viltain_01264R	CTGAAGGCGGACTC TTGATAGC	Dmel_Viltain_00536F	TCGCAGACATTGA AGCAGCC	729	60	60	Nathan Medd (Obbard lab) thesis
D.obscura Sigma virus	Sigmavirus (Rhabdoviridae) (-ssRNA)	(Longdon et *al*., 2010)	NC_022580	Dobs_SV_cons_F	YMGDCATTGGGGNC ATCC	Dobs_SV_cons_R	TCATCNGCCATNG TCAABCC	719	59	60	Longdon et al. 2010
Ashworth virus	Picornavirales (+ssRNA)	Obbard lab (Unpublished)	Awaiting submission	Ashw_A_2239F	CCAGGCAGAAACTTG TCA	Ashw_A_2901R	GCGCCAGTTTTAA GATCTC	662	54.5	60	Nathan Medd (Obbard lab) thesis
Sighthill virus	Bunyaviridae cf. Phleboviruses (-ssRNA)	Obbard lab (Unpublished)	Awaiting submission	DrosMix_Bunya_Chizelik e_02051F	GGATCCTTCTCATTA ACAGCC	DrosMix_Bunya_Chizel ike_02368R	GGTACTTCTTACTA ACGCTCC	317	55.8	60	Designed by M Wallace, Ch. 2 of this thesis
				Bunya_H_Tick_2like_010 64F	CTCCTGAATGTTTCT CCAAGG	Bunya_H_Tick_2like_0 1668R	GTCCTCAACTCCAT TAATCCC	604	56	60	
Vogrie virus	Reoviridae (dsRNA)	Obbard lab (Unpublished)	Awaiting submission	DrosMix_Reovirus1_0180 8F	CGACCCATATACATC AACCC	DrosMix_Reovirus1_02 617R	GATGGGTCTCGAA ATATCATCC	809	55.9	60	Designed by M Wallace, Ch. 2 of this thesis
				DrosMix_Reovirus2_0124 8F	ATGCCGGATCAATTC AGG	DrosMix_Reovirus2_01 769R	GTATGGGATATTC GCCACC	521	55	60	
Burdiehouse burn virus	Chuvirus (-ssRNA)	Obbard lab (Unpublished)	Awaiting submission	DrosMix_Chuvirus_02165 F	ATTCAGAACTCTCCT CTGATCC	DrosMix_Chuvirus_028 66R	CTGACAAGCCTTC TATGATGG	701	56.3	60	Designed by M Wallace, Ch. 2 of this thesis
Crammond virus	cf. Virgaviridae (+ssRNA)	Obbard lab (Unpublished)	Awaiting submission	Cramm_V_4951_F	TTCGCTGACTTTCTG ATGAAGA	Cramm_V_5856_R	GACTCCATGAGAA GCCCAAT	905	57.5	60	Designed by M Wallace, Ch. 2 of this thesis
Sunshine virus	Bunyaviridae cf. Phleboviruses (-ssRNA)	Obbard lab & Balinger Lab (Unpublished)	Awaiting submission	DrosMix_Bunyavirus_L4_ 00346F	CAAAGAAGCAAGAGT CCATCC	DrosMix_Bunyavirus_L 4_01828R	GCTAATTTCTCTAT TGCCCCC	1482	56.7	60	Designed by M Wallace, Ch. 2 of this thesis

Table S4. 1 PCR primers for scanning wild-collected flies. Details of primer assays and conditions used to scan pools of wild-collected D. immigrans, D. melanogaster and Obscura group species for viruses.

Virus	Forward primer name	Forward primer sequence (5'-3')	Reverse primer name	Reverse primer sequence (5'-3')	Product length (bp)	Tm (°C)
D. melanogaster Sigma virus	DMelSV_F_flap	AATAAATCATAATTCAATTTTGTACGCGGAATC	DMelSV_R_flap	AATAAATCATAATGATCAAACCGCTAGCTTCA	139	60
Torrey Pines virus	345_qPCR_1F_flap	AATAAATCATAATACATCAACGCTAACACGG	345_qPCR_1R_flap	AATAAATCATAATGGACGCATCAGTGAATAT	158	60
Craigies Hill virus	noda_qPCR_1F_flap	AATAAATCATAACGTTATGATTTATTCGTGGGCG	noda_qPCR_1R_flap	AATAAATCATAAGCGTCAAATATAATAGGTGCC	133	60
D. melanogaster Nora virus	Nora_qPCR_3F_flap	AATAAATCATAAGGTGTAGCAGGTCGTATTCTGC	Nora_qPCR_3R_flap	AATAAATCATAACAATGGCTGAAACTGCTGTTCCTGC	120	60
Chaq satellite	Chaq_514F_short	CGAAGTAACATACCAGCCATGG	222_660R	TCCATGTCCTGTHGGGTCTATCTG	126	57
	222_140F	AACAGAACGWCTGCTTTTTGGAAATCC	Chaq_263_R_sh	GCGCGGATCGTAAAATCGTCC	170	63
Galbut virus	Galbut_614F_sh	GTCCAGTAGTGAGAGATGCCACTC	407_750R	GCCKCATACTTGGTGCTGCCAACTG	165	63
	407_170F	GATCGAGATGGAACTCCRCTCTC	Galbut_293R_short	CTATTCCTAGTAGCCGGGAGTTC	124	60
Sighthill virus (L gene)	Si_hill_L_2231F_short	CAGATATGATTTGTGGAACTGC	DrosMix_Bunya_Chizelike_02368R	GGTACTTCTTACTAACGCTCC	137	55.5
Larkfield virus	DrosMix_Totivirus_01124F	GTACCAGATCATTGCTATGACC	Larkfield_1266_R_short	CTCTCACGGCGTCTTCG	142	57
Vogrie virus (seg 5)	Vog_seg5_1631F_short	GATACCATTCAATTCTCATCGC	DrosMix_Reovirus2_01769R	GTATGGGATATTCGCCACC	138	55
Brurdiehouse Burn virus	DrosMix_Chuvirus_02165F	ATTCAGAACTCTCCTCTGATCC	BraidieB_2314R_short	AAATGACAAGACGAGGATGC	149	56
Crammond virus	Cramm_V_5709F_short	GTCTGATCCTCACAGGTGG	Cramm_V_5856_R	GACTCCATGAGAAGCCCAAT	147	57
Buckhurst virus	Dobs_Buckhurst_05694F	GTGTCTTGGTGCGCACAAAT	Buckhurst_5852R_sh	GCACCTTAGAGACACATTAAGTTG	158	59

Withyham virus	Withyham_0821F_short	AGGATCATTCACATCAAGAATCTG	Dobs_Withyham_00970 R	ATGTGTCCGCCATCATCAAC	149	58
Pow burn virus	Pow_Burn_03794F_short	GACCATAAGAAAGCCAAGACTG	Dobs_PowBurn_B_03926R	TGTTAGCCTCTCCGTATGC	129	57
Grom virus	Dobs_Grom_00728F	CGGCAGGTCACAACATCTCAT	Grom_00873R_short	TTTGGTGGTGGTTGGTTGATC	145	60
Prestney burn virus	PrestneyB_00737F_short	CGGCAGGTCACAACATAATATCAG	Dsub_Prestney_Burn_A_00880R	TGGTGGTGGTTGGTTGATCG	143	60
Lye green virus	Dobs_Lye_Green_04753F	TTGTCGAGAATAGCAGGAGTCC	Lye_green_4943R_sh	GAAGAGAATGAAACGAGGTGGC	191	59.5
Eccles virus	Dsuz_Eccles_seg1_B_1664F	AGAAAGCAGGTTGATCGTCG	Eccles_seg1_13810R_short	AACTCAGGTTCATTGTGACCC	115	58
Blackford virus	Dtri_Blackford_01878F	ATCAAGCCGTCCGTGGAATC	Blackford_2011R_short	TGCTTGATCATGTAGCTGAACAC	133	60
Corseley virus	DsubDsuz_Corseley_00511F	ACGTGTTGAGCGAGGAGTAC	Corseley_00658R_sh	GCAATGCCAGAGTCCGTCTC	148	61
Bunyavirus L gene 4 / Sunshine virus	DrosMix_Bunyavirus_L4_00346F	CAAAGAAGCAAGAGTCCATCC	Buny_L4_479R_short	AAGCTCTAGGTCCACTTCC	133	56
D.immigrans Sigma virus	Dimm_SV_07812F_short	CAATGACAGGCCAGACATG	DImmSV_L_R2	GAATGCTCCAACCCTTTTGA	122	56
Muthill Virus	Muthill_6795F_sh	GCGTCACAAGTCCAAGTTCATG	Muthill_6943R_sh	GTGTGTGGGCTAGGAAACGC	148	61
D.imm Nora virus	Dimm_nora_2690F_sh	TCAAAGCAGGCATAATGAAGG	Dimm_nora_2825R_sh	TCAGTTTGAAGGAACTCATTGG	136	56.5

Table S4. 2 PCR primers for scanning female Dmel. Details of primer assays and conditions used to scan OreR Wol- females who were exposed to viruses carried by wild-collected flies. All assays used an extension time of 1 minute.

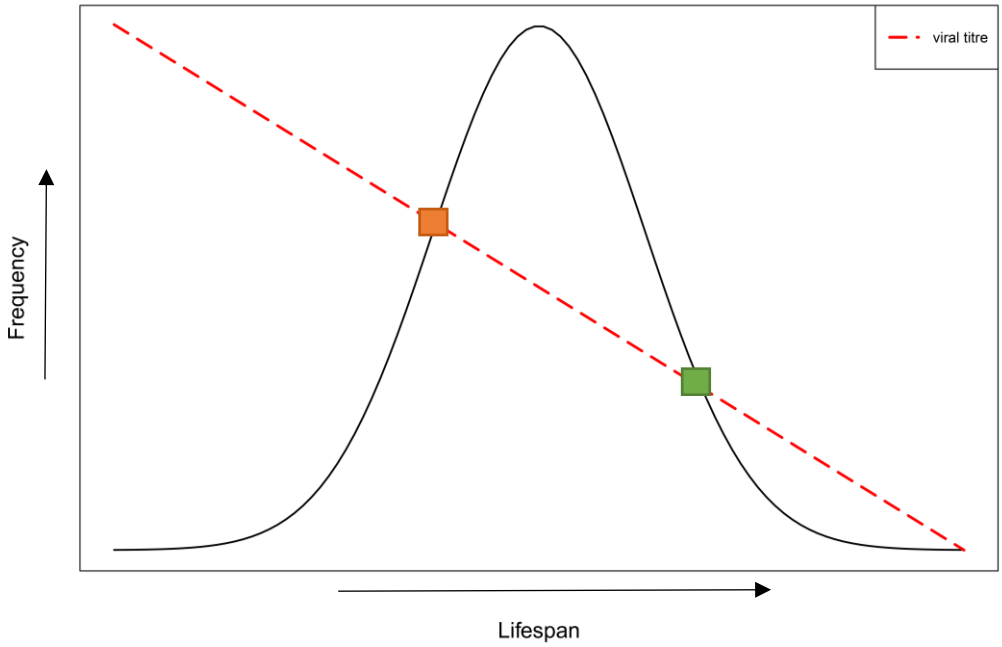


Fig. S4. 1 Potential issues with sampling lifespan with RNA virus clearance. The plot shows the potential sampling bias induced by sampling lifespan in a population where RNA viruses are being cleared by flies. If a fly dies earlier in the experiment (at the orange box), independent of any virus effect, there is a higher likelihood of detection of the virus because viral titre is higher. By comparison, if a fly dies later in the experiment (at the green box) there is less likelihood of virus detection. Therefore, lower lifespan observations might be falsely associated with presence of a virus, and higher ones with virus absence.

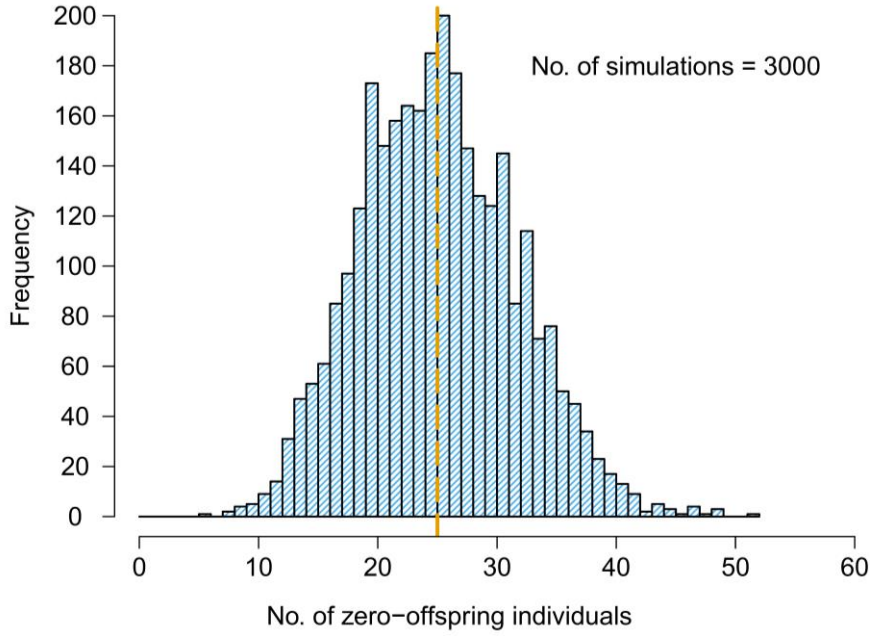


Fig. S4. 2. Predicted number of Dmel females producing zero offspring under the chosen hurdle model for lifetime offspring production. The frequency histogram shows the distribution of predicted zero-offspring values when the offspring data was simulated 3000 times under the hurdle model. The orange line indicates the actual number of zero-offspring individuals recorded in our data (25 females).

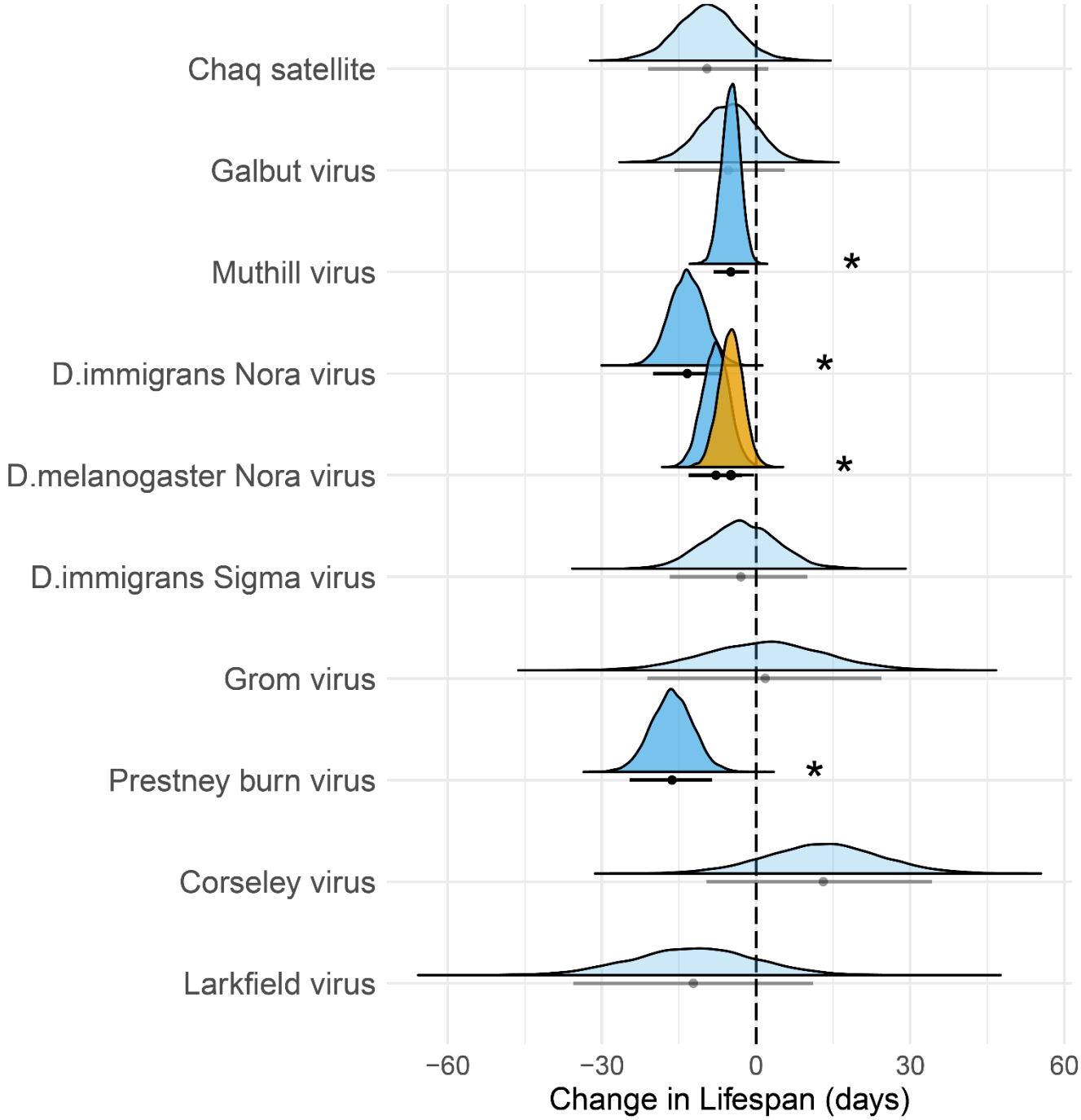


Fig. S4. 3 Posterior density distributions of the viral fixed effects included in the model of lifespan variation in Dmel females, with extra data for DmelNV. Solid horizontal lines and points indicate the 95% credible intervals, and posterior means, for each of the fixed effects included in the model.' Above these the distribution of the estimates is displayed as output from MCMCglmm. The original model distributions are displayed in blue, and the distribution of the effect of DmelNV infection when offspring and male infection data is included is shown in orange.

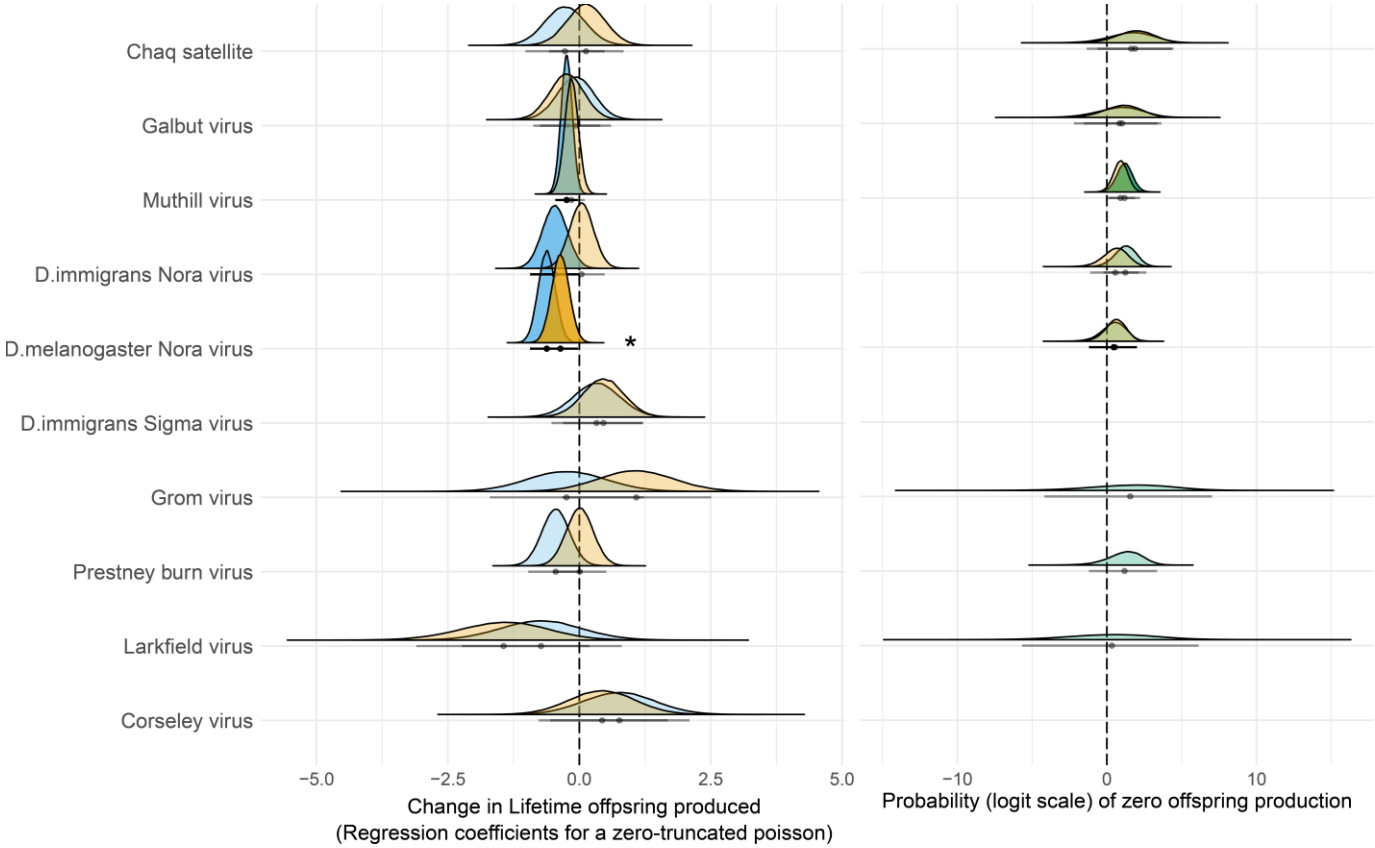


Fig. S4. 4 Posterior density distributions of the viral fixed effects included in the model of lifetime offspring production, and early life offspring production in Dmel females. Solid horizontal lines and points indicate the 95% credible intervals, and posterior means, for each of the fixed effects included in the model.' Above these the distribution of the estimates is displayed as output from MCMCglmm. Posterior distributions from the model of lifetime offspring production are displayed in blue (zero-truncated Poisson) and green (probability of zero offspring), and distributions from the model of early life offspring production in orange.

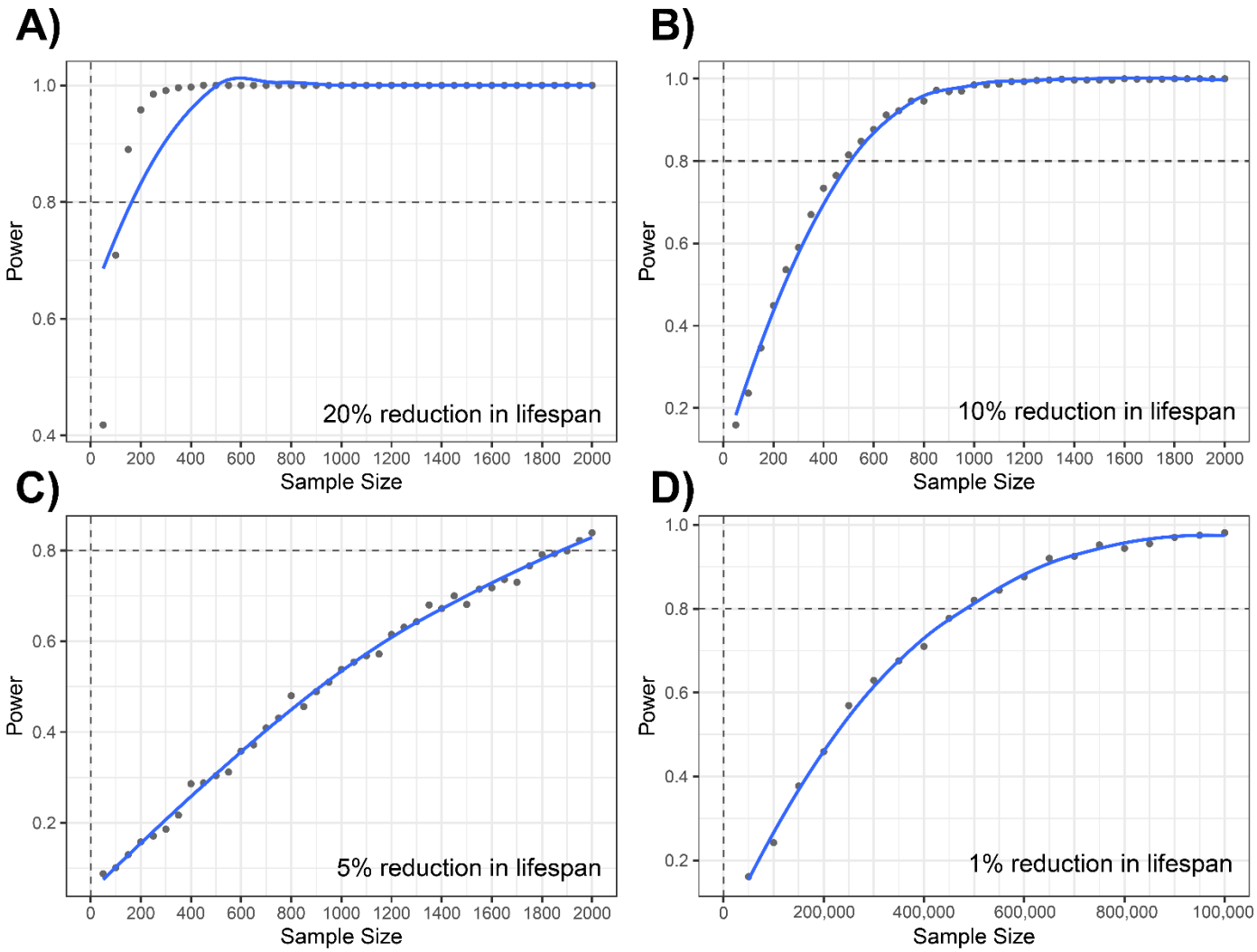


Fig. S4. 5 Simulation of power to detect a viral-induced reduction in lifespan. Plots show the sample size needed to detect a A) 20%, B) 10%, C) 5%, and D) 1% reduction in lifespan due to infection by a virus. Each data point represents the mean power (over 1000 simulations) to detect the reduction in lifespan, and how this changes as sample size is increased within each experimental group. The line indicates smoothed conditional means for these points. The simulations were created using the mean and variance in lifespan measured from this experiment.