THE UNIVERSITY
of EDINBURGH

The evolution and genetics of antiviral immunity in *Drosophila*

William H. Palmer

A thesis submitted for the degree of Doctor of Philosophy

The University of Edinburgh

2018

Contents

Declaration

The work within this thesis is my own, except when stated at the beginning of each chapter. Chapters 2-5 are either published or submitted manuscripts, and therefore I use “We”.

William H. Palmer

Abstract

Virus-host interactions determine virus transmissibility and virulence, and underlie coevolution that shapes interesting biological phenomena such as the genetic architecture of host resistance and host range. Characterization of the virus factors that exert selective pressure on the host, and the host genes which underlie resistance and adaptation against viruses will help to define the mechanistic pathways embroiled in host-virus coevolution. In this thesis, I describe the viral causes and host consequences of host-virus coevolution. These include genomic signatures consistent with antagonistic coevolution in antiviral RNA interference pathway genes such as high rates of positive selection and polymorphism, loci that underlie genetic variation in resistance to virus infection, and apparent conflict between NF-κB signalling and DNA virus infection.

The RNA interference (RNAi) pathway is the most general innate immune pathway in insects, underlined by the observation that many viruses encode suppressors of RNAi (VSRs). The relationship between RNAi and VSRs has garnered attention as a plausible battleground for host-virus antagonistic coevolution, and genomic patterns in *Drosophila* support this hypothesis. However, genomic patterns in the N-terminal domain of the key RNAi effector gene, Argonaute-2, have not been described. In Chapter 2, I sequence the Argonaute-2 N-terminal domain using PacBio long-read sequencing technology to describe variation within and across *Drosophila* species, and test whether this variation is associated with resistance to Drosophila C Virus.

The RNAi pathway evolves adaptively in *Drosophila*, but this has not been formally extended across invertebrate species. In Chapter 3, I quantify rates of adaptive protein evolution and describe evidence for selective sweeps in RNAi pathway genes using population genomic data from 8 insect and nematode species. These analyses indicate that RNAi genes involved in suppression of transposable elements and defence against viruses evolve rapidly across invertebrates, and I identify genes with signatures of elevated adaptation in multiple insect species.

Host genes that underlie host-virus interactions have been described in RNA virus infection of *Drosophila*, however substantially less attention has focussed on the host response to DNA viruses, primarily because no DNA viruses have been isolated from *Drosophila*. In Chapter 4, I describe the isolation of Kallithea virus, a *Drosophila* dsDNA nudivirus, and characterise the

host response to infection and genetic variation in resistance. I find that Kallithea virus infection causes early male-specific lethality, a cessation of oogenesis, and induction of undescribed virus-responsive genes. Further, I describe genetic variation in resistance and tolerance to Kallithea virus infection, and identify a potential causal variant for virus-induced mortality in *Cip4*.

Insect viruses commonly encode viral suppressors of RNAi, however there are a multitude of antiviral immune mechanisms besides RNAi which may select for viral-encoded inhibitors. In Chapter 5, I describe the requirement for RNAi and NF-κB in immunity against Kallithea virus, and map gp83 as a virus-encoded inhibitor of NF-κB signalling. I find that gp83 inhibits Toll signalling at the level of, or downstream of NF-κB transcription factors, and that this immunosuppressive function is conserved in other nudiviruses.

Lay Summary

Virus infection is detrimental to host survival and reproduction, and therefore hosts have evolved immune defences, which limit virus replication. Therefore, the immune system and viruses are in conflict, and if this relationship is sustained may enter an evolutionary arms race, where the host and virus genes involved are under continual evolutionary pressure to best each other. This can result in the maintenance of alleles that provide resistance to viruses, or rapid change in the host and virus genes involved, which together determine the range of host species a virus can infect, and how it spreads through populations. Therefore, identifying the host genes that evolve in response to viruses and how these genes change through time can give important insights into associations between hosts and viruses, and susceptibility to virus infection.

In this thesis, I characterise the evolution of the RNA interference pathway, an important part of the insect antiviral immune system. I find that the genes that are part of the RNA interference pathway generally have patterns consistent with evolution in response to viruses, and that this occurs broadly in invertebrate species. I also isolate a DNA virus (named Kallithea virus) that infects fruit flies, an important model organism useful for studying genetics and evolution. I find Kallithea virus is disadvantageous to fruit fly survival and reproduction and I identify the alleles in fruit fly populations that affect their ability to resist infection. Finally, because this type of virus has not been well-studied in insects, I describe the fruit fly genes involved in immune responses to Kallithea virus, and find a virus gene that suppresses these responses, indicating the genes involved may be a battleground where hosts and viruses co-evolve.

Acknowledgements

First and foremost, I would like to thank Darren Obbard for his enthusiastic guidance, realistic advice, and for always making himself approachable and engaging with problems or excitement. I would also like to thank Jarrod Hadfield for taking the time to give me statistical advice throughout my PhD. Also, many thanks to Ronald van Rij, for engaging with me about immune suppressors, supporting me during my time in his laboratory, and for his thoughts and advice on experiments and writing.

I would also like to thank all the people that made working in various laboratories so fun and productive, especially other Obbard lab members Fergal Waldron, Sam Lewis, Megan Wallace, and Nathan Medd, for great chats about (e.g.) snails, insect germlines, podcasts, and everything else, respectively. In the Ashworth fly lab, a special thanks to Pedro Vale and Craig Walling for supporting my sometimes-large use of fly food, and Helen Borthwick, Helen Cowen, and Angela Reid for making the food. Additionally, a huge thanks to Pedro and Katy Monteith for making sure the fly lab runs smoothly. In the Ashworth molecular lab, thanks to the members of the Vale, Ross, and Twyford labs for being good company and sharing reagents. Thanks to all the members of the van Rij lab for making Radboud a great place to work, and for helping me with finding my way around the lab and Nijmegen. Finally, thanks to Steve Mitchell for kindly helping with operation of the TEM.

Thanks to all those that funded my work or shared reagents. My PhD was supported by the Darwin Trust of Edinburgh, the sequencing in Chapter 2 was funded through a NERC NBAF, and my time at Radboud University was supported by an EMBO Short-Term fellowship. Thanks to Stuart McDonald, Maria-Carla Saleh, Marc Dionne, François Leulier, David Finnegan, and Bruno Lemaitre for sharing fly lines, and Pascale Dijkers, Jean-Luc Imler, Neal Silverman, and Edan Foley for sharing plasmids.

A personal thanks to all the people in Ashworth, especially those in my PhD cohort (±2 years), for being so fun, especially during the pub-ridden first two years. Finally, thanks to Ally Dillenburg, and my friends and family back home that visited me, for keeping me sane.

Chapter 1: General Introduction

The text of this chapter derives in part from my contribution to Palmer WH, Varghese FS, van Rij RP. Natural Variation in Resistance to Virus Infection in Dipteran Insects Viruses, 10.3 (2018) 118.

I am the sole author of all of the text used here, with comments on earlier drafts from Darren Obbard and Ronald van Rij. Finny Varghese created Figure 1.

1.1 Antiviral immunity in *Drosophila*

The *Drosophila* model has been integral in the description of virus-host interactions in insects, and the dissection of innate immune pathways with broad conservation across animals. Insect antiviral immunity is mediated by recognition and degradation of viral RNA (RNA interference), a transcriptional response orchestrated by signalling pathways such as Nuclear Factor κB and JAK-STAT, and cellular responses, such as apoptosis, phagocytosis, and autophagy (Figure 1).

1.1.1. RNA Interference

RNA silencing pathways utilize short RNA sequences bound to an Argonaute-family protein to transcriptionally or post-transcriptionally silence complementary "target" sequences. These include the microRNA (miRNA) pathway, which primarily regulates endogenous gene expression, the PIWI-interacting RNA (piRNA) pathway, which regulates transposons, and the small interfering RNA pathway (siRNA), which serves as one of the most important antiviral defences in invertebrates. Briefly, the siRNA pathway begins when Dicer-2 (Dcr-2), an RNase III family endonuclease, recognizes viral-derived dsRNA (which acts as a pathogen-associated molecular pattern – i.e. PAMP) and cuts it into 21 nucleotide siRNA duplexes (Lee *et al.*, 2004; Okamura *et al.*, 2004; Wang *et al.*, 2006; Deddouche *et al.*, 2008). A single strand of the siRNA duplex is loaded into the Argonaute-2 (Ago2)-containing RNAi Induced Silencing Complex (RISC), which then slices any viral sequence that is complementary to the loaded siRNA, thereby controlling both RNA and DNA virus infection (Kemp *et al.*, 2013). Recently, "secondary siRNAs" have been described, which are produced from virus-transposon DNA chimeras reverse-transcribed from defective viral particles (Goic *et al.*, 2013; Tassetto, Kunitomi and Andino, 2017; Poirier *et al.*, 2018). These secondary siRNAs are excreted in exosomes by he-

mocytes, and alongside the release of dsRNA from infected cells, mediate a non-cell autonomous, systemic RNAi-based immune response (Saleh *et al.*, 2006, 2009, Goic *et al.*, 2013, 2016; Tassetto, Kunitomi and Andino, 2017).

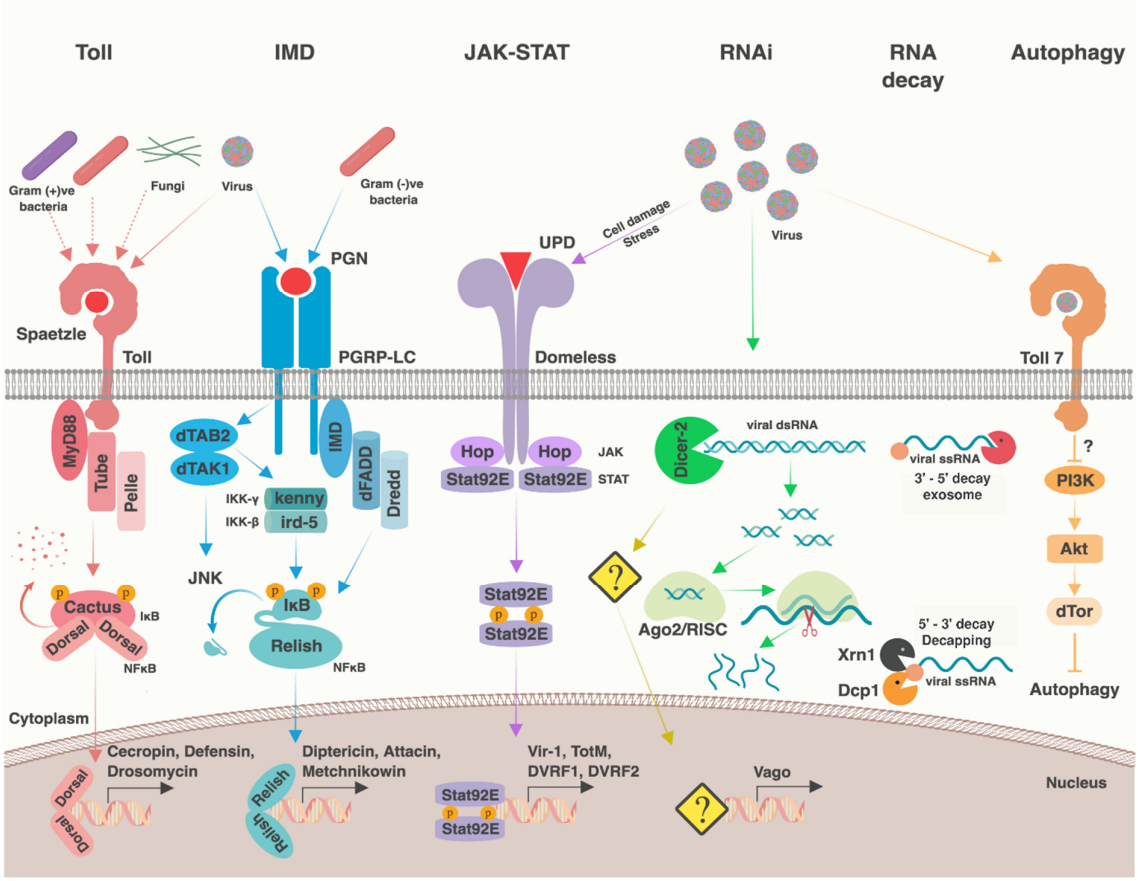

Figure 1. Overview of innate antiviral immune pathways in Drosophila.

The host pathways important during antiviral defence are listed above, including transcriptional responses orchestrated by Toll, Imd, and JAK-STAT, RNA-based immunity through RNA interference and decay, and cellular-based immunity such as autophagy. A simplified illustration of insect proteins that function in each pathway is shown below each pathway label. Figure created by Finny Varghese.

1.1.2. Other RNA Processes

In addition to RNAi, other RNA-based cellular processes have been implicated in antiviral defense in insects. The RNA decay pathway protects against defective cytoplasmic RNAs, including those without 5' caps or polyadenylated tails, lacking stop codons (nonstop decay), encoding early stop codons (nonsense-mediated decay), or stalled in ribosomes (no go decay) (Shoemaker and Green, 2012). Offending RNA molecules are deadenylated and either degraded 3' to 5' by the RNA exosome, or decapped and degraded by the 5' to 3' exonuclease

Xrn1 (Schoenberg and Maquat, 2012). Viral RNA may include a number of hallmarks of aberrant cellular transcripts, including 5' triphosphate groups, limited poly-A tails, or interior stop codons, making them susceptible to the RNA decay machinery (Moon and Wilusz, 2013; Molleston and Cherry, 2017). In *Drosophila*, the decapping enzymes, DCP1 and DCP2, and the RNA exosome have antiviral activity, and RISC-mediated silencing is partially dependent on Xrn1 and the exosome (Orban and Izaurralde, 2005; Hopkins *et al.*, 2013; Cao *et al.*, 2016; Molleston *et al.*, 2016). Additionally, there is evidence of viral modulation of processing body (P body) and stress granule formation. For example, Cricket paralysis virus (CrPV) infection leads to dispersal of P body components, which are sites that may be associated with organized decapping and 5' to 3' decay (Khong and Jan, 2011). In mosquitoes, evidence for antiviral RNA decay comes from the flavivirus noncoding RNA, sfRNA, which is produced by stalling of Xrn1 on structured RNA elements in the 3' UTR (Moon *et al.*, 2012).

RNA editing by adenosine deaminases that act on RNA (ADAR) occurs during the mammalian innate immune response, where the replacement of adenosines with inosines reduces the stability of dsRNA and the accuracy of replication and translation, due to inosines being read as guanosines (Samuel, 2012). Evidence in support of an antiviral activity of ADAR in insects comes mainly from mutation and substitution patterns in a subset of RNA viruses. Two sequenced strains of sigma virus (DMelSV) appeared to have been hypermutated from A to G, and analysis of DMelSV variation within flies and across populations found that ADAR-resistant sites are less likely to be polymorphic (Carpenter *et al.*, 2009; Piontkivska *et al.*, 2016).

1.1.3 Nuclear Factor κB Pathways

There are two Nuclear factor κB (NF-κB) signalling cascades in flies: the Toll and immune deficiency (Imd) pathways (Valanne, Wang and Ramet, 2011; Myllymaki, Valanne and Ramet, 2014). Generally, these pathways are activated when an upstream pattern recognition receptor interacts with a PAMP. The PAMPs that trigger NF-κB activation in response to bacteria and fungi include conserved cell wall components, although the viral PAMPs responsible for pathway activation are unknown. PAMP recognition leads to phosphorylation of IκB (inhibitor of κB; encoded by cactus for the Toll signalling and by the C-terminal domain of relish in Imd signalling) by an IκB kinase (IKK, encoded by pelle in Toll signalling and kenny in Imd signalling). This results in degradation of IκB and subsequent release of the NF-κB transcription factor (encoded by dorsal or Dorsal-related immune factor for Toll signalling and relish for Imd signalling). These transcription factors then translocate into the nucleus to induce

transcription of immune effectors, including antimicrobial peptides (Valanne, Wang and Ramet, 2011; Myllymaki, Valanne and Ramet, 2014). Notably, activation of Imd may also signal through the JNK pathway (Silverman *et al.*, 2003; Delaney *et al.*, 2006). Although the primary functions of these pathways are in anti-bacterial and anti-fungal defense, both have been implicated in defense against various dipteran viruses (Zambon *et al.*, 2005; Xi, Ramirez and Dimopoulos, 2008; Avadhanula *et al.*, 2009; Costa *et al.*, 2009; Kemp *et al.*, 2013; Ferreira *et al.*, 2014), and consistent with this, some insect viruses encode suppressors of the Imd pathway (Lamiable, Kellenberger, *et al.*, 2016).

The antiviral effectors downstream of NF-κB are, for the most part, unknown, besides a couple of examples. The NF-κB-responsive antimicrobial peptides are slightly upregulated after viral challenge, and some have antiviral properties (Luplertlop *et al.*, 2011; Huang *et al.*, 2013). However, the antiviral activity of the Toll and IMD pathways may also be based on mechanisms independent from antimicrobial peptides. For example, Toll signalling is involved in differentiation of hemocytes – phagocytic cells in the hemolymph of invertebrates (Qiu, Pan and Govind, 1998). Imd signalling can be pro-apoptotic, which itself can have antiviral functions (Georgel *et al.*, 2001; Liu *et al.*, 2013), and induces Pvf2 expression which activates antiviral ERK signalling (Xu *et al.*, 2013; Sansone *et al.*, 2015).

1.1.4. JAK-STAT Pathway

In insects, the JAK-STAT pathway is activated upon binding of one of the unpaired ligands (upd, upd2, or upd3) to the pathway receptor, domeless. This interaction results in activation of the JAK kinase hopscotch, which then phosphorylates Stat92E, resulting in its dimerization and translocation into the nucleus where it induces transcription of JAK-STAT dependent genes (Arbouzova and Zeidler, 2006). The antiviral gene Vago, which is induced downstream of dsRNA recognition by Dcr-2, may also activate JAK-STAT signalling in mosquitoes, although possibly through a different receptor (Deddouche *et al.*, 2008; Paradkar *et al.*, 2012; Kemp *et al.*, 2013; Jupatanakul *et al.*, 2017). During infection with some viruses, *Vago* and the *upd* ligands are upregulated, as well as a subset of known STAT-regulated genes (Dostert *et al.*, 2005; Merkling *et al.*, 2015; West and Silverman, 2018). This activation results in a transcriptional program which can mediate infection tolerance or antiviral resistance, depending on the virus (Dostert *et al.*, 2005; Souza-Neto, Sim and Dimopoulos, 2009; Paradkar *et al.*, 2012; Kemp *et al.*, 2013; Jupatanakul *et al.*, 2017). Similar to antiviral NF-κB signalling, there are

virus-specific differences in STAT-responsive transcriptional output, and STAT-responsive antiviral effectors are still mostly unknown. However, *attC* is upregulated downstream of JAK-STAT following Sindbis virus (SINV) infection in Drosophila, and heterozygous *attC* mutations lead to increased viral replication (Huang *et al.*, 2013).

1.1.5. Apoptosis, Phagocytosis, and Autophagy

Apoptosis, phagocytosis, and autophagy make important contributions to dipteran antiviral defense. In *Drosophila*, viral infection induces apoptosis through p53-mediated transcription of RHG genes (*reaper, hid, grim*, and *sickle*), which promote degradation of Drosophila inhibitor of apoptosis 1 (DIAP1) and consequent activation of the initiator caspase Dronc (Steller, 2008; Liu *et al.*, 2013). This response occurs rapidly after infection, effectively reducing the duration a virus can access host factors that are crucial for replication (Liu *et al.*, 2013). Additionally, an N-terminal degron (a protein domain that regulates protein turnover rate) renders DIAP1 inherently unstable (Vandergaast *et al.*, 2015), allowing the promotion of apoptosis during virus-mediated translational inhibition (Vandergaast *et al.*, 2011). Phosphatidylserine on apoptotic cells is recognized by the engulfment receptors draper and Integrin βv on phagocytes, which eliminate infected cells (Nainu *et al.*, 2015). Individual viral particles, including Drosophila C Virus (DCV) and White spot syndrome virus (a natural shrimp nimavirus) may also be direct targets of phagocytosis in cell culture (Ye and Zhang, 2013; Zhu and Zhang, 2013). Finally, autophagy, a process in which intracellular particles are enveloped by membrane crescents and shuttled to lysosomes for degradation, is also antiviral in some contexts (Shelly *et al.*, 2009; Moy and Cherry, 2013; Lamiable, Arnold, *et al.*, 2016). For example, noncanonical Toll-7 signalling may be responsible for activating antiviral autophagy during rhabdovirus and bunyavirus infection, although autophagy is dispensable against most other viruses (Nakamoto *et al.*, 2012; Moy *et al.*, 2014; Lamiable, Arnold, *et al.*, 2016).

1.2 Host-virus coevolution in *Drosophila*

The fitness of hosts and viruses are intricately linked, because an increase in the fitness of one often corresponds to a decrease in the fitness of the other. If this antagonistic relationship is sustained, hosts and viruses are expected to co-evolve, with each adapting to overcome the others' defences and counter-defences (Brockhurst *et al.*, 2014). This may drive the evolution of viruses, host immune genes, and host genes hijacked by viruses for replication,

and can shape the genetic variation of host resistance in populations. For example, *D. melanogaster* harbours substantially more genetic variation in resistance to its native viruses than artificially introduced 'non-natural' infections (Magwire *et al.*, 2012). Two primary models have been put forward to describe co-evolutionary dynamics: the 'arms race' model, whereby new adaptive mutations arise and are fixed in hosts and pathogens, and the negative frequency dependent selection model, where rare host alleles are more fit but lose their selective benefit as they increase in frequency (Ebert, 2008). These are expected to leave behind distinct genomic patterns, including increased divergence or polymorphism, respectively, in the genes underlying the conflict. However, host–virus interactions are immensely complex (Sessions *et al.*, 2009), and most of the loci under pathogen-mediated selection are unknown. Identification of these loci is of great medical and evolutionary importance, and efforts have been made to characterize genetic variation in host resistance, and to determine the effects of parasite-mediated selection on host gene evolution. In the *Drosophila* model, genomic signatures of parasite-mediated selection have been investigated in genes segregating antiviral resistance polymorphisms and genes in antiviral innate immune pathways.

1.2.1 Viral resistance loci and their evolution

In *D. melanogaster*, the most extensively studied host–virus system at the population level is DmelSV, a negative-sense ssRNA rhabdovirus (Longdon and Jiggins, 2012). DmelSV is transmitted vertically to offspring through eggs and sperm, although some fly strains are not permissive to DmelSV replication and transmission (Fleuriet, 1976). This resistance was first mapped to a complex amino acid substitution in the N-terminal PB1 domain of *ref(2)P* (Table 2), a gene now known to be involved in autophagy (Fleuriet, 1976; Wayne, Contamine and Kreitman, 1996; Avila *et al.*, 2002; Carré-Mlouka *et al.*, 2007). Viral replication and transmission is reduced in homozygous *ref(2)P* mutants or trans-heterozygotes bearing a mutant *ref(2)P* and the refractory allele, indicating *ref(2)P* is a proviral host factor in DmelSV infection (Contamine, Petitjean and Ashburner, 1989). Subsequent mapping experiments identified the *CHKov* paralogues and *Ge-1* as loci with segregating variants associated with resistance to DmelSV. The first is a triallelic polymorphism at the *CHKov1* and *CHKov2* paralogs, where an insertion of a transposable element in *CHKov1* provides resistance to DmelSV, and a following complex (and rare) rearrangement is associated with even greater DmelSV resistance (Magwire *et al.*, 2011). Although the functional relevance of the *CHKov* genes during viral infection remains untested, the resistance mutations could exert an effect on viral entry, as

CHKov is a predicted acetylcholine esterase, and Rabies virus uses the acytelcholine receptor as a cell entry point (Lentz *et al.*, 1983). A large deletion in the serine-rich linker region of *Ge-1* results in a 10-fold reduction in viral titer and lower infection rates (Gay, 1978; Cao *et al.*, 2016). Ge-1 is a conserved adaptor bridge between Decapping protein 1 and Decapping protein 2 that helps localize these enzymes to processing bodies (Yu *et al.*, 2005; Jinek *et al.*, 2008), and thus could plausibly exert its antiviral effect through the RNA decay pathway (Cao *et al.*, 2016). Finally, up to 7 QTLs associated with DmelSV replication or transmission have been identified in genetic mapping experiments, however the exact loci responsible have yet to be reported (Gay, 1978; Brun and Plus, 1980; Bangham *et al.*, 2008; Cogni *et al.*, 2016).

Natural variation has also been mapped for Drosophila C Virus (DCV), a ssRNA positive sense dicistrovirus. Resistance to DCV has been mapped to complex polymorphism at the *pastrel* (*pst*) locus (Magwire *et al.*, 2012; Martins *et al.*, 2014). The primary *pst* resistance mutation is due to a single amino acid change, although there are multiple structural alleles and cis-regulatory changes that may enhance resistance, resulting in seven alleles with four distinct phenotypes (Cao *et al.*, 2017). The function of pst is unknown, but overexpression of the susceptible allele provides protection against DCV, and *pst* is upregulated after intra-abdominal viral injection, indicating that *pst* is an induced antiviral factor (Cao *et al.*, 2017). In addition to *pst*, polymorphism in *Anaphase promoting complex 7*, *Ubiquitin conjugating enzyme E2H*, and 2 QTLs may also underlie genetic variation in DCV resistance (Magwire *et al.*, 2012; Martins *et al.*, 2014).

The genetic architecture of resistance to viruses in *Drosophila* often seems to include large-effect polymorphisms at an intermediate frequency in populations. These variants are likely under pathogen-mediated selection, and surrounding patterns of polymorphism and divergence have been compared to expected patterns of balancing or recurrent directional selection, which would be compatible with their involvement in antagonistic coevolution (Anderson and May, 1982; Ebert, 2008; Obbard and Dudas, 2014). Although none of the identified resistance loci display significantly increased diversity that is hallmark of balancing selection, *ref(2)P* and *pastrel* have relatively high levels of nonsynonymous or structural polymorphism (Dru *et al.*, 1993; Wayne, Contamine and Kreitman, 1996; Magwire *et al.*, 2012; Cao *et al.*, 2017). This may be due, in part, to incomplete selective sweeps, whereby a resistance mutation rises to high frequency and subsequently loses the selective benefit through viral counter-adaptation. This was likely witnessed in the 1980s, when a DmelSV

strain that regained transmission advantage in flies with a resistant *ref(2)P* allele swept through the population (Fleuriet, 1986; Fleuriet and Sperlich, 1992; Fleuriet and Periquet, 1993; Bangham *et al.*, 2007; Wilfert and Jiggins, 2013).

In addition to maintaining polymorphism, pathogen-mediated selection may result in fixation of adaptive mutations. This can occur quickly on an evolutionary timescale, and these beneficial mutations will only be briefly visible as variation within populations. However, they can be recognized as elevated divergence between populations or species, or by the diversity patterns they leave behind. This is evident in genes with segregating resistance polymorphism, and *ref(2)P*, *CHKov1/2*, and *Ge-1* show signs of recent or recurrent positive selection (Bangham *et al.*, 2007; Magwire *et al.*, 2011; Cao *et al.*, 2016).

1.2.2 Evolutionary patterns of *Drosophila* immune genes

Immunity genes are likely targets of pathogen-mediated selection, and some display signatures of long-term balancing selection or recurrent positive selection. For example, nonsynonymous polymorphism in NF-κB-responsive AMPs appear to be under balancing selection, with some convergently maintained in different *Drosophila* species, likely due to their role in anti-bacterial or anti-fungal immunity (Unckless and Lazzaro, 2016; Unckless, Howick and Lazzaro, 2016). Also, the N-terminal region of Ago-2 is hypervariable (Hain *et al.*, 2010), either due to diversifying selection or a combination of high mutation rate and drift.

Pathogen recognition proteins, such as rel and its interactors, and RNAi genes show elevated rates of adaptive protein evolution (Schlenke and Begun, 2003; Sackton *et al.*, 2007; Juneja and Lazzaro, 2010; Han *et al.*, 2013). In *Drosophila melanogaster*, RNAi genes mediating defence against transposons and viruses, but not those involved in control of endogenous transcripts (i.e. the miRNA machinery), have increased rates of adaptive evolution, and *Ago2*, *Dcr-2,* and *R2D2* are among the fastest evolving genes in the *D. melanogaster* genome. Additionally, six *Ago-2* loci from four *Drosophila* species (*D. melanogaster*, *D. simulans, D. yakuba,* and 3 paralogues in *D. pseudoobscura*) exhibit diversity patterns consistent with positive selection driving low frequency mutations to fixation, such as a local reduction of diversity and linked high-frequency derived polymorphisms (Kolaczkowski, Hupalo and Kern, 2011; Obbard *et al.*, 2011; Lewis *et al.*, 2016).

1.2.3 Host immune evolution driven by immune suppressors

The patterns of positive selection found in *Drosophila* RNAi genes are especially intriguing because of the prevalence of virus-encoded suppressors of RNAi (VSRs) in insect-infecting viruses. VSRs have evolved independently in different viral families and inhibit RNAi through disparate mechanisms, including interference with RISC assembly, shielding viral dsRNA from Dcr2, or marking RNAi pathway components for degradation (Li, Li and Ding, 2002; Van Rij *et al.*, 2006; Nayak *et al.*, 2010; van Mierlo *et al.*, 2012; Bronkhorst *et al.*, 2014; Van Cleef *et al.*, 2014). Regardless of mechanism, RNAi-VSR conflict and resulting coevolution is an enticing hypothesis to explain adaptation in RNAi genes. In support of this, the *D. immigrans* Nora virus VSR, VP1, inhibits AGO2 in its native host species (i.e. *D. immigrans*), but not in *D. melanogaster*, indicating VP1 has evolved specificity for inhibiting the *D. immigrans* RNAi pathway (van Mierlo *et al.*, 2014). Additionally, VSRs tend to have higher rates of protein evolution than other viral genes, consistent with the idea that host resistance and viral counter-resistance may be continually evolving in 'arms race' dynamics (Obbard *et al.*, 2009).

1.3 DNA viruses of Drosophila

Most studies on antiviral immunity and host-virus coevolution in *Drosophila* have used RNA viruses largely because, until recently, there were no described *Drosophila* DNA viruses (reviewed in Xu and Cherry, 2014). Some of these have been *Drosophila*-infecting viruses, used to explore natural host-virus relationships, and include sigma viruses (Rhabdoviridae), DCV (Dicistroviridae), Drosophila A Virus (Picornavirales), and the Nora viruses (Picornavirales) (Brun and Plus, 1980; Ambrose *et al.*, 2009; Habayeb, Ekström and Hultmark, 2009; Longdon, Obbard and Jiggins, 2010; van Mierlo *et al.*, 2014). These viruses, alongside additional viruses from disparate families and hosts, have been used to dissect the antiviral immune mechanisms, and explore the diversity of virus-host interactions. For example, Flock House Virus (from a beetle; Nodaviridae), Drosophila X Virus (natural host unknown, Birnaviridae), and Cricket Paralysis Virus (Dicistroviridae) are all virulent in flies, and therefore fly mortality and virus titre can be jointly assayed to study host mechanisms of resistance and tolerance (e.g. Settles and Friesen, 2008; Martins *et al.*, 2014; Merkling *et al.*, 2015; Gupta and Vale, 2017). *Drosophila* is also susceptible to many vertebrate-infecting, mosquito-borne viruses, including alphaviruses, bunyaviruses, rhabdoviruses, reoviruses, and flaviviruses (Xu and Cherry, 2014), making *Drosophila* a valuable model for studying the insect interactions with medically relevant viruses. For example, genome-wide RNAi screens have identified host

factors impacting West Nile and dengue virus infections, some of which are conserved in the mosquito and vertebrate hosts (Sessions *et al.*, 2009; Yasunaga *et al.*, 2014).

In contrast to the diversity of RNA viruses studied in *Drosophila*, fly anti-DNA viral immunity has only been studied using IIV6 – a moth iridovirus. Experiments with IIV6 have shown that the RNAi and JAK-STAT pathways can provide resistance and tolerance to DNA viruses, respectively (Bronkhorst *et al.*, 2012; Bronkhorst *et al.*, 2014; West and Silverman, 2018). However, because *D. melanogaster* and IIV6 are not naturally associated in the wild, IIV6 cannot be used to study coevolutionary processes, including evolutionary outcomes such as naturally segregating resistance polymorphisms in the host, or host-specific adaptations in the virus. Instead, addressing these questions requires a native *Drosophila* DNA virus isolate.

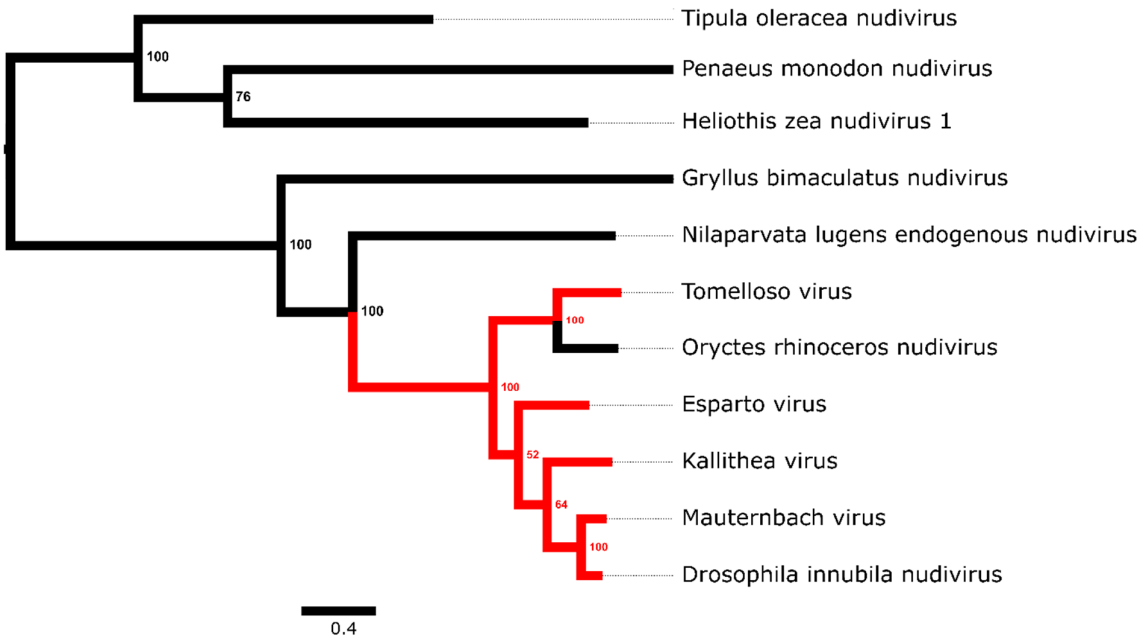

Figure 2: Nudivirus phylogeny

A phylogeny of the polymerase B gene from all sequenced nudiviruses. The tree was created in PhyML from a protein alignment, with an LG substitution model, allowing substitution rate to vary according to a discrete gamma distribution with four rate categories. Branches in red are Drosophila-associated viruses.

Unckless (2011) described the first DNA virus of *Drosophila innubila* (Drosophila innubila nudivirus; DiNV) – a geographically widespread nudivirus that causes lethality and loss of fecundity in infected flies. Since the description of DiNV, metagenomic sequencing has led to

the discovery of four more *Drosophila* nudiviruses: Esparto virus, Tomelloso virus, Mauternbach virus, and Kallithea virus (red in Figure 2) (Webster *et al.*, 2015; Kapun *et al.*, 2018). Of these, Kallithea virus is the most widespread, present in *D. melanogaster* populations worldwide at a global prevalence of approximately 5%.

1.3.1 Nudivirus replication

Nudiviruses are large enveloped rod-shaped dsDNA viruses with approximately 100 genes and 200 kilobase (kb) genomes. Nudiviruses enter cells either through endocytosis or pinocytosis, and the nucleocapsid is released into the cytoplasm following fusion between the viral envelope and host endosomes (Crawford and Sheehan, 1985; Fay and Panté, 2015). Although transport and nuclear entry of nudiviruses has not been investigated, baculovirus (a closely related viral family) nucleocapsids induce polymerization of actin, which shuttle the virus to the nucleus, and enter the nucleus through the nuclear pore complex (Fay and Panté, 2015). Transcription begins as early as 2 hours post-infection, and early cellular signs of DNA replication of Oryctes nudivirus (OrNV) occurs roughly 7 hours post-infection (hpi) in cell culture, after which chromatin-free areas are observed in the nucleus, and nucleocapsid assembly begins (Crawford and Sheehan, 1985). A similar pattern of infection is observed for Heliothis zea Nudivirus 1 (HzNV-1), with only slightly longer incubation times (Ralston, Huang and Kawanishi, 1981). During early OrNV infection, double membrane vesicles can be observed throughout the nucleus of infected midgut and fat body cells, and are subsequently filled by nucleocapsids, either through their recruitment, or assembly within the membrane shells (Huger, 1966; Payne, 1974; Crawford and Sheehan, 1985). In some cases, long filamentous structures are observed in the nucleus, which are thought to be assembly lines of nucleocapsids (Huger, 1966). During late infection, nucleocapsids form densely packed lines at the nuclear membrane, at which stage the nuclear membrane appears disrupted, the nucleus is hypertrophied, and the cytoplasm can become disorganised (Huger, 1966; Payne, 1974). Replication peaks at 12 hpi for OrNV and 24 hpi for HzNV-1, after which viral particles can be observed budding from the cytoplasm, or released from the nucleus following lysis of the nuclear and plasma membranes (Ralston, Huang and Kawanishi, 1981; Crawford and Sheehan, 1985).

1.3.2 Nudivirus transmission and pathogenesis

Four nudiviruses have been studied *in vivo* in insects: *Drosophila innubila nudivirus* (DiNV), Gryllus bimaculatus nudivirus (GbNV), Oryctes rhinoceros nudivirus (OrNV), and Heliothis zea Nudivirus 2 (HzNV-2). DiNV, GbNV, and OrNV are likely transmitted faecal-orally. Infections with GbNV and OrNV can be achieved through infected food, and DiNV and OrNV particles are found in the faeces of infected flies and beetles, respectively (Huger, 1966, 1985, Zelazny, 1973a, 1976; Boucias, Maruniak and Pendland, 1989; Unckless, 2011). Although the pathology has been best studied in OrNV infection, GbNV and DiNV infections suffer from increased mortality and in the case of DiNV, decreased fecundity, demonstrating obvious fitness costs associated with infection and indicating there is likely undescribed pathogenesis occurring (Huger, 1985; Boucias, Maruniak and Pendland, 1989; Unckless, 2011). OrNV is highly virulent, especially for infected larvae, and has been used as a biocontrol agent to protect coconut palms from *Oryctes* beetles in the South Pacific (Alois M. Huger, 2005). In larvae, OrNV infects the midgut and fat body, causing hypertrophied nuclei, vacuolization and necrosis of infected fat body cells, and a prolapsed hindgut (Huger, 1966; Alois M. Huger, 2005). As a result, larvae stop feeding, become lethargic, and eventually die. OrNV is also transmitted among adult beetles, where infection of midgut cells induces proliferation of surrounding apical cells. These continue to replicate until the gut lumen is filled with infected cells, likely interfering with nutrient absorption, and could be linked with observed reduced fecundity, flight activity, and lifespan in infected beetles (Zelazny, 1973b, 1977; Alois M. Huger, 2005).

Compared with the other nudiviruses, HzNV-2 has a remarkably different transmission strategy and pathology. HzNV-2 infection can occur following ingestion by larvae, but it primarily replicates in the adult gonads and is transmitted sexually among adult moths and transovarially to offspring (Raina and Adams, 1995). Both sexes suffer from severe gonadal atrophy following infection, and a 'waxy plug' of viral particles forms in the oviduct, ostensibly assisting sexual transmission (Raina and Adams, 1995; Hamm, Carpenter and Styer, 1996).

1.3.3 Nudivirus genes

There are 12 sequenced nudivirus species, all of which have been found in invertebrates. These nudiviruses encode between 98 and 160 open reading frames (ORFs), with 32 genes shared across nudiviruses and 20 shared with baculoviruses (Wang and Jehle, 2009). These include 9 loci involved in DNA replication and nucleotide metabolism, 6 genes involved in

transcription (including the late expression factor, or "lef" genes), 4 peroral infectivity factors ("pif" genes, required specifically for oral infection), and 4 genes required for packaging and assembly (Wang and Jehle, 2009). In OrNV, 13 proteins are associated with the nucleocapsid, and 14 proteins with the envelope, although these have not been identified (Crawford and Sheehan, 1985). Consistent with this, HzNV-1 has 28 proteins associated with both envelope and nucleocapsid, although GbNV only has 17 (Burand, Stiles and Wood, 1983; Boucias, Maruniak and Pendland, 1989). Most nudiviruses also encode an inhibitor of apoptosis, however this gene has been lost along the lineage to Esparto virus, Mauternbach virus, Kallithea virus, and DiNV. The remaining genes have no described function, and approximately half lack any recognisable protein domains.

HzNV-1 and KV both encode at least one miRNA, although they do not share sequence similarities. The HzNV-1 gene *hhi1* is highly expressed during early infection, promoting replication and causing apoptosis in infected cells (Wu, Wu, Liu, Hsu, *et al.*, 2011). During late infection, the HzNV-1 miRNA becomes the only highly expressed locus, and induces a latent stage of infection, at least partially through the downregulation *hhi1* (Wu, Wu, Liu, Lee, *et al.*, 2011). KV encodes 2 putative miRNAs, one of which is highly expressed in wild infected flies, although the function is unknown (Webster *et al.*, 2015).

The evolution of nudivirus genes has recently been assessed through the comparison of OrNV, KV, and DiNV (Hill and Unckless, 2018). Gene order is highly conserved between the closely related DiNV and KV, however breaks down between more distantly related viruses (Wang and Jehle, 2009; Hill and Unckless, 2017a). Individual genes evolve mostly under purifying selection, although *helicase* and *ODV-E56* genes may be under reduced constraint or positive selection (Hill and Unckless, 2018). The latter is supported by an analysis of baculovirus genes, which identified *helicase* as a target of positive selection (Hill and Unckless, 2017b).

1.4 Thesis Aims

In this thesis, I characterise patterns of host-virus coevolution in insect RNAi genes, isolate Kallithea virus – a *Drosophila* DNA virus, and characterise host-virus interactions between *Drosophila melanogaster* and Kallithea virus, focussing on products of host-virus conflict such as host resistance polymorphisms and viral immune evasion strategies. In Chapter 2, I describe variation in and evolution of a rapidly evolving repetitive region of Ago2, and assess

whether observed repeat sequence variation is associated with viral resistance. In Chapter 3, I extend previous observations of adaptive divergence in *Drosophila* RNAi genes across invertebrate species, and quantify genic rates of adaptive evolution to identify pervasive positive selection in RNAi pathways in conflict with viruses and transposable elements. Because the known insect antiviral repertoire is biased towards anti-RNA viral defences, I then isolate Kallithea virus, and describe viral resistance polymorphisms and transcriptional responses to infection. Finally, through a small-scale screen for immune suppressors, I identify *gp83* as a conserved nudivirus-encoded Toll pathway inhibitor, suggesting Toll may be important at the immune interface in *Drosophila*-DNA virus conflicts.

Chapter 2: Variation and evolution of the Glutamine-rich repeat region of Argonaute-2

The text of this chapter is published:

Palmer, William H., and Darren J. Obbard. "Variation and evolution in the glutamine-rich repeat region of Drosophila Argonaute-2." G3: Genes, Genomes, Genetics 6.8 (2016): 2563-2572.

I wrote this chapter with comments and minor textual edits from Darren Obbard. Thanks to Francis Jiggins and Daniel Fabian for making the DGRP virus survival data available to us.

2.1 Abstract

RNA interference pathways mediate biological processes through Argonaute-family proteins, which bind small RNAs as guides to silence complementary target nucleic acids. In insects and crustaceans *Argonaute-2* silences viral nucleic acids, and therefore acts as a primary effector of innate antiviral immunity**.** Although the function of the major *Argonaute-2* domains, which are conserved across most Argonaute-family proteins, are known, many invertebrate *Argonaute-2* homologs contain a glutamine-rich repeat (GRR) region of unknown function at the N-terminus. Here we combine long-read amplicon sequencing of Drosophila Genetic Reference Panel (DGRP) lines with publicly available sequence data from many insect species to show that this region evolves extremely rapidly and is hypervariable within species. We identify distinct GRR haplotype groups in *D. melanogaster*, and suggest that one of these haplotype groups has recently risen to high frequency in a North American population. Finally, we use published data from genome-wide association studies of viral resistance in *D. melanogaster* to test whether GRR haplotypes are associated with survival after virus challenge. We find a marginally significant association with survival after challenge with Drosophila C Virus in the DGRP, but we were unable to replicate this finding using lines from the Drosophila Synthetic Population Resource panel.

2.2 Introduction

Argonaute proteins are the effectors of eukaryotic RNA interference (RNAi) pathways, using short nucleic acid guide sequences to target complementary sequences for transcriptional or post-transcriptional repression. RNAi-related pathways mediate a diverse range of biological

processes, from regulation of developmental genes through miRNAs and endogenous siRNAs, to defence against genomic parasites such as transposable elements via piRNAs (Carmell *et al.*, 2002; Meister, 2013). In insects, antiviral RNAi is mediated by an RNA Induced Silencing Complex that contains Agonaute-2 (Ago2). This complex is guided by 21nt siRNAs 'diced' from viral replicative intermediates and other dsRNA substrates by Dicer-2 (Lee *et al.*, 2004; Okamura *et al.*, 2004; Wang *et al.*, 2006) and bound to Ago2. Ago2 then uses these siRNAs to target the 'slicing' of viral single-stranded RNA, rendering the targeted viral genome or transcript non-functional.

Despite the diverse biological roles played by Argonaute proteins, their structural organisation is generally conserved over deep evolutionary time (Swarts *et al.*, 2014). For example, eukaryotic Argonaute proteins have a PIWI domain that binds and/or 'slices' target nucleic acids (Parker, Roe and Barford, 2004; Song *et al.*, 2004), MID and PAZ domains that bind the 3' and 5' ends of the small RNA, respectively (Lingel *et al.*, 2003; Ma, Ye and Patel, 2004; Ma *et al.*, 2005; Boland *et al.*, 2010), and an N-domain which is involved in duplex unwinding (Kwak and Tomari, 2012). Nevertheless, in contrast to these highly conserved domains, the N-terminal region of Argonaute proteins tends to be disordered and lack sequence complexity, and is highly variable between species (Hain *et al.*, 2010). This variation is particularly striking in the arthropod antiviral gene, Ago2, where the N-terminal region is often composed of numerous glutamine-rich repeat motifs ('GRR') (Hain *et al.*, 2010). For example, even between closely related species such as *Drosophila melanogaster* and *D. simulans*, the N-terminal sequence divergence is extensive. In *D. melanogaster*, Ago2 includes one of the most repetitive amino acid sequences in the genome (Jorda and Kajava, 2009), while in *D. simulans* it is markedly different, with only one large duplication of almost the entire N-terminus (Figure 1, Figure 2).

In *D. melanogaster,* the GRR region is composed of two distinct repeat regions (GRR1 and GRR2) (Hain *et al.*, 2010). The most N-terminal, GRR1, is a 6 amino acid imperfect repeat (QQLQQP) present in two to four copies, while GRR2 is a 23 residue imperfect repeat (Figure 2) previously reported to occur between seven and eleven times in succession in laboratory strains (Hain *et al.*, 2010). Although genetic studies have elucidated the function of Ago2 in *D. melanogaster,* the role of the GRR is still unknown. In other proteins, long poly-glutamine rich regions have been implicated in increased protein adhesion and protein complex for-

mation, and underlie numerous human diseases (Fan *et al.*, 2014). However these are generally long contiguous tracts of glutamine residues, in contrast to the short complex repeat units observed in the Ago2 GRR. Further, *Ago2* GRR deletions appear to have no effect on RISC assembly in *Drosophila* (Liu *et al.*, 2009), suggesting that this domain is not required for binding siRNAs or catalysing target cleavage.

The absence of known function makes it difficult to predict which evolutionary forces underlie the observed rapid evolution of the GRR. In contrast, and consistent with the antiviral role of *Drosophila* Ago2, the other domains of this protein display strong evidence of positive selection: they exhibit locally reduced diversity around the gene through selective sweeps, and elevated rates of amino acid substitution (Obbard *et al.*, 2006, 2011; Kolaczkowski, Hupalo and Kern, 2011). We have previously argued that this rapid adaptive evolution may be driven by virus-mediated selection, through the action of viral suppressors of RNAi (Obbard et al., 2009), such as those seen in Drosophila C Virus and *Drosophila melanogaster* Nora Virus (Van Rij *et al.*, 2006; van Mierlo *et al.*, 2014). The reportedly high level of variation within the *D. melanogaster* GRR region is therefore surprising, as diversity is expected to be continually removed by nearby selective sweeps. One possible explanation is that the high diversity and differentiation seen in the GRR is purely a result of low constraint on this sequence, combined with high rates of recombination and replication slippage mediated mutations (Jeffreys *et al.*, 1988). Alternatively, if the GRR domains are involved in the antiviral function of Ago2, or interact with VSRs, the high diversity seen in Ago2 GRRs could reflect the action of diversifying selection—which is a common outcome of many models of host-parasite coevolution (Antonovics and Thrall, 1994; Sasaki, 2000).

Whether or not the high divergence and diversity seen in GRR2 is an evolutionary consequence of virus-mediated selection, a virus-related role for GRR2 might be reflected by segregating functional variation associated with GRR2 haplotype. In principle, this could be identified by a genome-wide association study (GWAS) such as that which identified *pastrel* (Magwire *et al.*, 2012). However, as repeat variants are challenging to reconstruct or identify using short sequencing reads (Treangen and Salzberg, 2012), GWAS analyses have largely been limited to SNP and simple structural variation. Thus, previous GWAS analyses of viral resistance in *Drosophila* (Magwire *et al.*, 2011, 2012) have been unable to test for phenotypes associated with highly repetitive sequences, and instead could only have detected its impact through linkage with neighbouring SNPs. But, because the SNP diversity is low in the

region surrounding *Ago2*, the scale of linkage disequilibrium (LD) is short in *Drosophila*, and the LD between a SNP and neighbouring hypermutable loci breaks down rapidly (Sawaya, Jones and Keller, 2016), a role for GRR variation in determining viral resistance remains untested.

Here we characterise the sequence diversity of the Ago2 GRR region in insects, and use Pacific -Biosciences SMRT long-read sequencing of RT-PCR amplicons to generate full GRR haplotypes for 127 lines of the *Drosophila* Genetic Reference Panel (DGRP) (Mackay *et al.*, 2012). We use these data to re-examine the evolution of this domain and its potential role in antiviral defence. In doing so we not only demonstrate the value of long-read technology for performing genome-wide association studies (GWAS) when complex repetitive loci are present, but also illustrate the potential challenges associated with such analysis using short-read technology alone. We provide the first robust *Ago2* GRR haplotypes for natural populations, identify likely haplotypes in publicly available short read data, and quantify differences in the frequency and composition of GRR haplotypes between African and North American populations. Using published GWAS data (Magwire *et al.*, 2012) to test for an association between GRR haplotype and virus survival phenotypes, we detect a small but nominally significant association of GRR haplotype with longevity of DCV-infected flies. However, we were unable to confirm this association with a second independent experiment using recombinant inbred lines.

2.3 Methods

2.3.1 Comparison of the GRR across insects

We obtained the GRR repeat unit for other insect species by using tBLASTx with default parameters to query all arthropod RefSeq RNA sequences using the *Ago2* region just C-terminal to the GRR from *D. melanogaster*. We manually selected repetitive sequences as input for Tandem Repeat Finder (v4.07b) (Benson, 1999) with a mismatch and indel penalty of 5 and minimum alignment score of 50. The insect reference tree was inferred using MrBayes (v2.13) (Huelsenbeck and Ronquist, 2001) with an HKY85 substitution model and gamma-distributed rate variation with invariable sites, using conserved sequences from the original tBLASTx search aligned in MUSCLE (v3.8.31) (Edgar, 2004) as input. The high divergence between GRR sequences, including extensive indel variation, makes it extremely challenging to infer positional homology (i.e. alignment) in the GRR regions (see figure 2 for *Drosophila*

alignments). We therefore used the frequency feature profile phylogeny building tool (v.3.19) (Sims *et al.*, 2009) to quantify similarity between the GRR of insects, as this approach can be used in the absence of alignment. Frequency feature profiles break the nucleotide or amino acid sequence into a distribution of kmers and compares these distributions against each other taking into account similarity between amino acid residues. The frequency feature profiles were constructed in two ways: in the first, GRR repeat unit consensus sequences were used as input to cluster GRRs, and in the second the entire GRR region was used. In each case, the topology of these clusters were compared to the MrBayes tree, using a kmer size which maximised similarity of the feature frequency profile tree to the MrBayes tree, as it is expected that the GRR shares the same history as the rest of Ago2.

2.3.2 Sample preparation

We sequenced the GRR region from a subset of the *Drosophila* Genetic Reference Panel (DGRP) and 7 other closely related *Drosophila* species. The DGRP constitute a collection of highly inbred lines from *D. melanogaster* collected in Raleigh, NC in 2003 (Mackay *et al.*, 2012) that have previously been sequenced using the Illumina platform to provide a public resource for GWAS. However, as short-read sequencing cannot easily be used to reconstruct repetitive sequences such as the GRR region of *Ago2*, we generated new amplicon data for the *Ago2* GRR region from 127 of these lines. To avoid sequencing the long intron between GRR1 and GRR2, (RT-)PCR was performed on RNA extracted from 10 flies per line to obtain an amplicon containing the full *Ago2* GRR1 and GRR2 regions. For *Drosophila* species other than *D. melanogaster*, sample origins are as described in Longdon et al (2011). For all species, RNA was extracted using Trizol (Ambion) according to the manufacturer's instructions. Three forward primers were designed separately for the *Drosophila melanogaster/simulans/mauritiana* clade, the *Drosophila yakuba/erecta/santomea* clade, and for *D. ananassae* based on published genome sequences (PCR primer sequences: 15F *D. yakuba*: ATGGGAAAGAAGAACAAATTCAAGG; 30F *D. melanogaster*: GAACAAGAAAGGAGGACAGG; 18F *D. ananassae*: ATATAAGGATGACGGGAAGC). PCRs shared a single reverse primer designed to amplify all species (1550R CAGCTTATCCACCGAGTAGCA) except for *D. ananassae* (GTCGACATTAAGAAACGGTT). Paired barcode sequences from the Pacific Biosciences SMRT Portal v1.4 were added to the 5' end of each primer, along with the padding sequence GGTAG. Barcoded amplicons were then combined into 10 pools of 16 samples and gel purified for sequencing.

2.3.3 Long read amplicon sequencing and analyses

Samples were pooled in groups of 16 and subject to Pacific Biosciences SMRT-cell sequencing (NERC Biomolecular Analysis Facility, Liverpool). *D. melanogaster* raw reads were demultiplexed and filtered in the SMRT portal by 5 minimum passes around the circular template, requiring a minimum predicted accuracy of 70%, a minimum insert size of 1000 bases, and a minimum barcode score of 22. From these, 5-pass circular consensus sequences (5CCS) were called for each read (raw read processing was performed by NERC Biomolecular Analysis Facility, Liverpool). Although these 5CCS reads may still contain errors, to obtain the final consensus sequence for each fly line we grouped all 5CCS reads by length, and then removed reads whose length was observed in less than 10 reads. This filtering resulted in a single peak of read lengths for each amplicon (e.g. Figure S1) in all but one fly line. In this one line (DGRP-306), we detected two high-frequency haplotypes, suggesting that this line is heterozygous at the GRR region, and this sample was excluded from all subsequent analyses. Consensus sequences from 5CCS reads within the length class resulted in high-confidence haplotypes from 127 of the DGRP lines, which were used in further analyses. In addition, eight haplotypes from a Kenyan (Nairobi) population, which were previously obtained by Sanger sequencing of long PCR products (Obbard *et al.*, 2006), were also included in the analysis. GRR sequences were also obtained from single lines of *D. simulans, D. sechellia, D. yakuba, D. santomea, D. erecta*, and *D. ananassae*. These were analysed in the SMRTportal with the parameters described above, but with a minimum insert size of 500 bp. We used BLAST (2.2.31+) (Camacho *et al.*, 2009) to recover the species of each long 5CCS read from the coding sequence to the 3' of GRR2, then we grouped reads by species and read length. Peaks in read length were again assumed to be indicative of a distinct amplicon, and analyses were performed as in the *D. melanogaster* samples. To cluster haplotypes (Figure 3) by repeat unit, the distinct repeat units observed in *D. melanogaster* were each labelled with an identifying letter, such that a haplotype can be denoted a string of repeat-unit identifier letters. We then used text-based feature frequency profiles (hash length of 2) to cluster and visualise haplotypes by repeat unit similarity (Sims *et al.*, 2009).

2.3.4 Characterisation of GRR repeats in published short-read data

To explore the utility of published short-read sequencing in the reconstruction of the *Ago2* GRR, we obtained short read sequences of DGRP (Accession number: PRJNA36679) and *Drosophila* 'Nexus' lines (Lack *et al.*, 2016) (Table S1 Accession numbers). To retain reads deriving

from the region of interest, all reads were mapped to our full set of 127 sequenced GRR haplotypes using Bowtie2 (v. 2.2.4) (Langmead and Salzberg, 2012) with default parameters, retaining all read pairs for which at least one mate mapped. An attempt was made to assemble these reads de novo using Velvet (v1.2.10) (Zerbino and Birney, 2008), using the hash length for each individual that maximised contig length, and using the expected coverage and insert length data provided by the sequence read archive.

To assess whether the distribution of repeat units in short-read sequences could be used to infer GRR2 haplotypes, we used Jellyfish (v.2.2.3) (Marçais and Kingsford, 2011) with a kmer size of 69 (the size of a GRR repeat in *D. melanogaster)* and a lower coverage bound of 2 (although this parameter had no qualitative effects when varied from 0 to 10) to infer counts for known repeat units in each sample. To ensure we only included samples with sufficient coverage of the GRR to reliably infer haplotypes, we filtered out those samples without reads supporting repeat unit GRR2-G and repeat unit GRR2-A, and without ten reads supporting GRR2-E (these repeat units were shown to occur in all 127 DGRP samples using PacBio amplicon sequencing, with GRR2-E being most common). The retained samples were then normalised by total read count to obtain a proxy for relative abundance of repeat units in each sample.

2.3.5 Linkage Disequilibrium analysis

We combined our GRR haplotype data with known SNPs and indels within 5 KB on either side of *Ago2* from the DGRP dataset (http://dgrp2.gnets.ncsu.edu/data/website/dgrp2.tgeno), replacing any reported sequence within the GRR with our own long-read sequence data. We then calculated a multiallelic extension of r^2 (Hill and Robertson, 1968), which provides an accurate metric of linkage disequilibrium (LD) among multiallelic loci (Zhao, Fernando and Dekkers, 2007). The analysis was performed using our data coded either as entire haplotypes (and therefore highly multiallelic), or as a series of SNPs and indels from alignment of the haplotypes.

The rapid increase in frequency of a beneficial allele is expected to lead to extended regions of high LD around the swept allele (termed 'haplotype homozygosity') (Sabeti *et al.*, 2002) and to quantify this, we used the program nSL (v.0.47) (Ferrer-Admetlla *et al.*, 2014), to calculate the nSL statistic for the regions surrounding the GRR. The nSL statistic is similar to the more widely-used iHS statistic (Voight *et al.*, 2006), except that distance is measured

as the number of segregating sites rather than map distance, making it more robust to recombination rate variation. Moving along the sequence, at each polymorphic site nSL calculates the average number of consecutive polymorphisms associated with either the ancestral or derived allele in question. Either exceptionally large or small values of the nSL statistic are evidence that a variant has rapidly increased in frequency. For *D. melanogaster* we polarised the sites with the *D. simulans* genome by parsimony, aligned by LastZ (v.1.02.00), and standardised the nSL statistic by allele frequency.

2.3.6 Association with viral phenotypes and infections

To test whether variation in the GRR haplotype is associated with variation in viral resistance, we used data from previous GWAS studies (Magwire *et al.*, 2011, 2012) of the DGRP lines for resistance against three different viruses. These were Drosophila C Virus (DCV, a horizontally transmitted and highly pathogenic dicistrovirus naturally infecting *D. melanogaster*); D. melanogaster Sigma Virus (DMelSV: a vertically transmitted Rhabdovirus naturally infecting *Dmel*), and Flock House Virus (FHV, a horizontally transmitted Alphanodavirus naturally infecting beetles, closely related to Newington virus of *D. immigrans* (Brun and Plus, 1980; Longdon and Jiggins, 2012; Webster *et al.*, 2015, 2016). We fitted general linear mixed models using the R package MCMCglmm (v2.22, Hadfield, 2010) with DGRP line and replicate block (block equivalent to date for FHV and DCV) as random effects, and known segregating functional variants (*pastrel* for *DCV*, and *ref(2)p, CHKov,* and *ge1* for DMelSV) and GRR haplotypes as fixed effects.

The final model was:

$$Y_{ijkl} \sim \mu + pastrel_i + haplotype_j + line_k + block_l + \epsilon$$

Where μ is the mean survival time and ϵ is a normally-distributed error term. If linkage disequilibrium is sufficiently large, it may be difficult to separate the effect of GRR haplotype from the effect of (partially) linked SNPs. Therefore, to examine whether the GRR haplotype is acting as a marker for a neighbouring causal SNP, we also fitted models in which each flanking SNP was tested for an association with mortality, and then selected those which were nominally significant (with no correction for multiple testing) for inclusion in the model outlined above, to verify any observed effect was due to the GRR.

2.3.7 Recombinant inbred line infections

To further test for an association between *Ago2* GRR haplotype and viral resistance, we experimentally infected recombinant inbred lines from the *Drosophila* Synthetic Population Resource (King, Macdonald and Long, 2012) with DCV. We categorised lines by *Ago2* GRR haplotype groups based on presence of reads containing the repeat units GRR2-L (as a marker for haplotype group alpha) or GRR2-D and GRR2-K (as markers for haplotype group beta) in the short-read data for the DSPR parental lines. The length of the linked region around the GRR region was calculated in each recombinant inbred line, and 100 lines from each haplotype group were selected with the aim of minimising the impact of linked variants (i.e. lines were chosen on the basis of nearby break points). Infections were performed by injecting *Drosophila C Virus* abdominally into 10 flies per vial with an average of 3 vials per line, at 10^5 TCID50, chosen on the basis that this dosage caused mortality in approximately one week. Flies were kept at 25 degrees in agar vials and monitored for 7 days post-infection (DPI) with mortality recorded on each day. The data were analysed using a binomial regression in MCMCglmm with the model:

$$Y_{hijklm} \sim \mu + DPI_h + DPI_h^2 + pastrel_i + haplotype_j + line_k + vial_l + date_m + date{:}\,DPI_{mh} + vial{:}\,DPI_{lh} + \epsilon$$

We followed Longdon et al (2011) in coding mortality (Y) as a number of 'successes' (the number of flies remaining alive in a vial on a certain day) and 'failures' (the number of flies that died on a certain day). This model fits GRR genotype, *pastrel* parent of origin (as a proxy for *pastrel* genotype), and DPI as fixed effects. DPI is encoded as both a linear and quadratic predictor, as mortality tends to decrease after the peak infection. We included DSPR line (genetic background), vial, and date as random effects, allowing for interactions between the DPI and either date or vial effects.

2.4 Results

2.4.1 Evolution of the GRR across insects

The presence of a GRR region in *Ago2* is conserved across the arthropods, but the GRR evolves extremely rapidly, and the diverse structure of the GRR makes alignment and assembly of these regions challenging. Some species have multiple repeat units, such as *Megachile rotundata* (leafcutter bee)—with repeat units QRRSLAPHG and LKQQQQPLAPQQHHTFA—others have nested repeat units, as in *Tribolium castaneum* (flour beetle), where a region

with multiple repeats with consensus QQQWQQQQPQPHP appears to have been duplicated. To circumvent the challenge of alignment difficulties, feature frequency profiles (distance matrices produced by comparing the distribution of kmers across different sequence) of the GRR and amino acid composition were used to quantify similarity without alignment. Conservation of either amino acid composition or repeat unit sequence could imply functional significance of the GRR, and so we examined the GRR of 34 insect species (Figure 1). Trees from feature frequency profiles were constructed from the entire GRR (Figure 1B) or from the consensus repeat unit (Figure S2), and compared to the Ago2 gene tree (Figure 1A, Figure S3). In both cases, the GRR region sequences clustered broadly according to known species relationships but do not reliably reflect more divergent evolutionary relationships. For example, the relationships between *D. melanogaster, D. simulans, D. erecta,* and *D. yakuba* were correctly resolved, but the Drosophilidae did not cluster together in any distance measure (for alignments, see Figure 2). This divergence is in part due to structural differences between GRRs (Figure 1C), as the number and size of repeat units is variable, even between closely related species. In addition, trees made from repeat unit consensus sequences are unable to correctly cluster hymenopterans and dipterans, indicating the divergence is unlikely to be due to assembly artefacts. Alternatively, amino acid sequence composition is similar across the species analysed, with glutamine the most frequent amino acid residue in all species analysed except *Athalia rosae* (turnip sawfly; Figure 1). This conservation is further illustrated throughout the Drosophilidae (and closest outgroup *M. domestica*), whose GRR is strikingly glycine-rich. These observations argue that although the GRR sequence and structure evolves quickly, the composition may be under selective constraint, implying functionality.

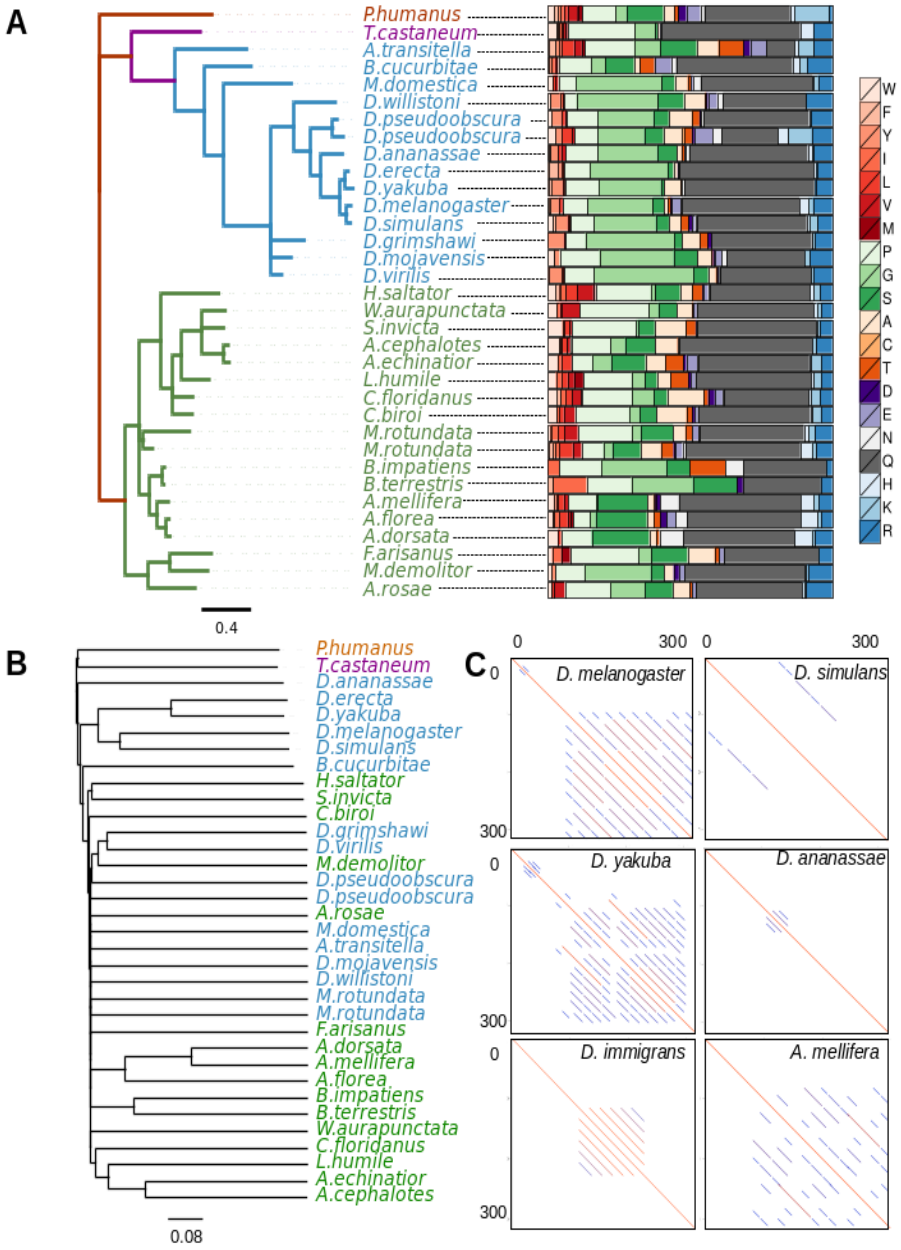

Figure 1: GRR evolves rapidly but maintains similar sequence composition

(A) The gene tree of conserved Ago2 sequence C-terminal to the GRR, for selected insect species, along with the corresponding amino acid residue composition of the entire GRR for that species. Hymenopteran species are coloured green and dipteran species are coloured blue. Across the insects analysed there is conservation of the residues from which the GRR is composed. Amino acid colour scheme: WFYILVM (reds, hydrophobic), PGS (greens, small secondary structure breakers), ACT (oranges, small amino acids), DE (purples, larger amino acids), NQ (greys, carboxamide side chains), and HKR (blues, electrically charged side chains). (B) Neighbour joining tree drawn from FFP clusters derived from the protein sequence of the entire GRR region: the lack of internal resolution reflects the rapid divergence of the GRR among species. (C) The GRR structure can change rapidly among closely related species. Shown are dotplots for the N-terminal 300 amino acids of Ago2 (plotted against itself) in D. melanogaster, D. simulans, D. yakuba, D. ananassae, D. immigrans, and A. mellifera. In these dot-plots the diagonal line from corner to corner represents the sequence identity to itself, and the successively shorter parallel lines reflect the multiple scales of self-similarity within the sequence (see Figure 2 for alignments).

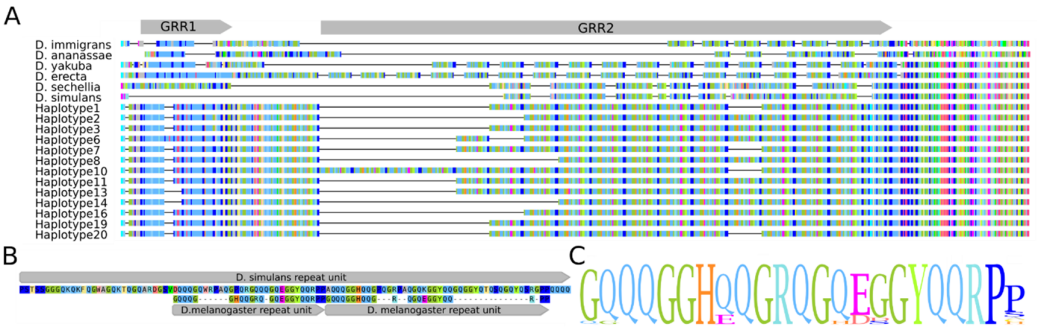

Figure 2: Drosophila GRR alignments

(A) Alignments of the GRR among seven Drosophila species for which the complete N-terminal region is available, including a subset of the newly sequenced DGRP haplotypes. Grey arrows above the alignment denote the GRR1 and GRR2 regions that align poorly across species, with the start of the more conserved Ago2 sequence to the right. (B) Alignment between D. simulans and D. melanogaster GRR repeat units showing how the D. simulans repeat motif appears to derive from a pair of neighbouring motifs in D. melanogaster, and exemplifying the rapid protein evolution of this region. (C) A sequence logo built from the alignment of unique repeat units found in D. melanogaster. The total height reflects overall diversity, and the height of the letter at each position signifies the frequency of that amino acid across distinct repeat motifs.

2.4.2 Haplotypes and repeat units in *D. melanogaster Ago2 GRR*

We found extensive repeat polymorphism among the DGRP lines. Among the 127 lines sequenced, we identified three different GRR1 haplotypes and eighteen GRR2 haplotypes, between which there is no detectable linkage disequilibrium (Figure S4). GRR1 and GRR2 regions could be identified in other *Drosophila* species we sequenced as well, however the repeat unit sequences differ, as described above (Figure 2). All GRR1 haplotypes comprise one to three perfect repeats of the sequence PQQLQQ, with two repeats being most common (Figure 3). The GRR2 is more complex, with 12 different repeat units (labelled GRR2-A to GRR2-L, Figure 3). The distinct repeat units seen in *D. melanogaster* are all within 3 nucleotide differences of each other and a consensus sequence of GQQQGGHQQGRQGQEGGYQQRPP (Figure 2), and occur 10-15 times in tandem. Most of the GRR2 sequence is composed of two repeat units: GRR2-E (occurring 4-8 times per haplotype) and GRR2-G (occurring 1-6 times per haplotype), which differ at a single amino acid position. In contrast, the majority of repeat units are rare—only occurring in one haplotype, and are most likely the result of recent single base-pair mutations (e.g. GRR2-J). Together, the GRR1 and GRR2 alleles form 23 distinct GRR haplotypes in our dataset. Clustering GRR haplotypes by repeat unit composition (see materials and methods) identifies two largely distinct haplotypes classes (coloured gold and red in

Figure 3), and one putatively recombinant haplotype (*GRR Hap11* – coloured orange) in the DGRP sample. Based on this clustering dendrogram, we have attempted to reconstruct the recent history of the GRR region, as most haplotypes appear to differ from one another by one or two mutation or recombination events (single base changes, whole-repeat insertions and/or deletions, and gene conversion) (Figure 4).

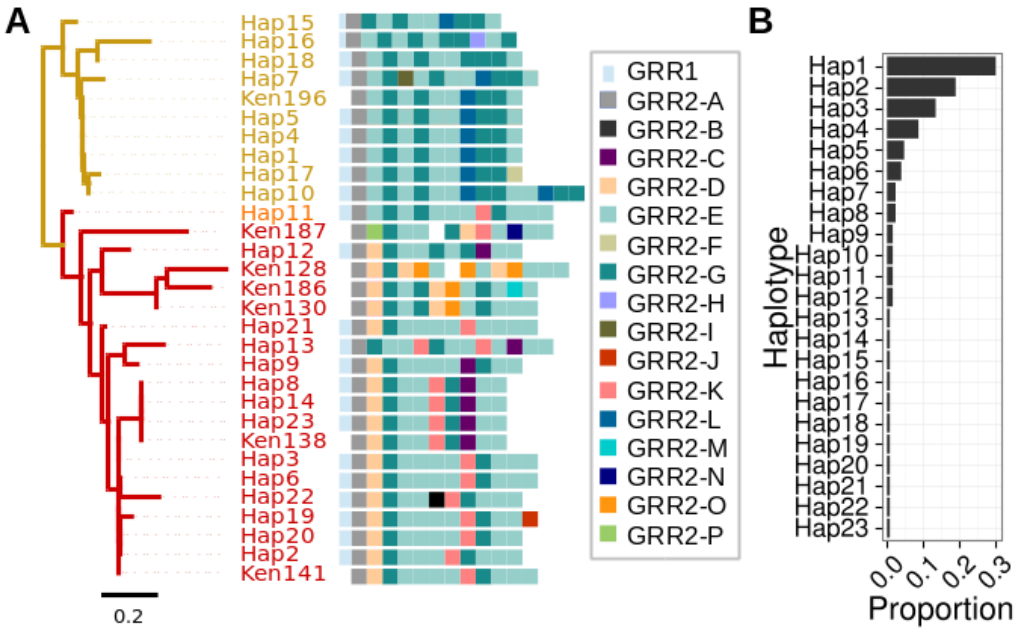

Figure 3: Variation in the GRR repeat sequence and structure

(A) Similarity clustering analysis of GRR2 haplotypes reveals two large groups of D. melanogaster haplotypes (gold and red) and one putatively recombinant haplotype (orange). Haplotypes are illustrated using colour-codes for the 16 distinct repeat units corresponding to the arbitrary character identifiers A-P. In some Sanger-sequenced Kenyan haplotypes (labelled 'Ken') a repeat unit could not be determined, denoted by a white square. Note that repeat unit L is diagnostic of haplotype group alpha, and units D and K are diagnostic of group beta. (B) Histogram of the frequency of each haplotype in the DGRP population. Most haplotypes occur at low frequency, with some high and intermediate frequency haplotypes.

Many of these GRR haplotypes occur at a low frequency in the DGRP, with 11 of the 23 haplotypes occurring only once in our sample (Figure 3, Figure 4). There are three high frequency haplotypes (*Haps1 – 3*) with the latter two differing by only one repeat unit. Interestingly, there are many differences between the *Hap1* and *Hap2/Hap3* groups (hereafter referred to as haplotype groups alpha and beta), such that no simple single mutational event could convert one to the other. Further, the haplotypes in haplotype group alpha occur at low frequencies and are no more than two mutational events from *Hap1* itself, suggesting they may have been formed recently. This observation is at odds with the high frequency of *Hap1*, and may

indicate a recent increase in the frequency of the GRR alpha group. In support of this idea, despite the approximately equal frequency of individuals in haplotype groups alpha and beta, nucleotide diversity in Ago2 and a 100 kb surrounding region is much lower in haplotypes from the alpha clade than those in the beta clade, indicating this *GRR Hap1* is younger than expected given its relative frequency (Figure S5). Nevertheless, there does not seem to be any evidence for significant extended haplotype homozygosity in the remainder of the gene (Figure S6).

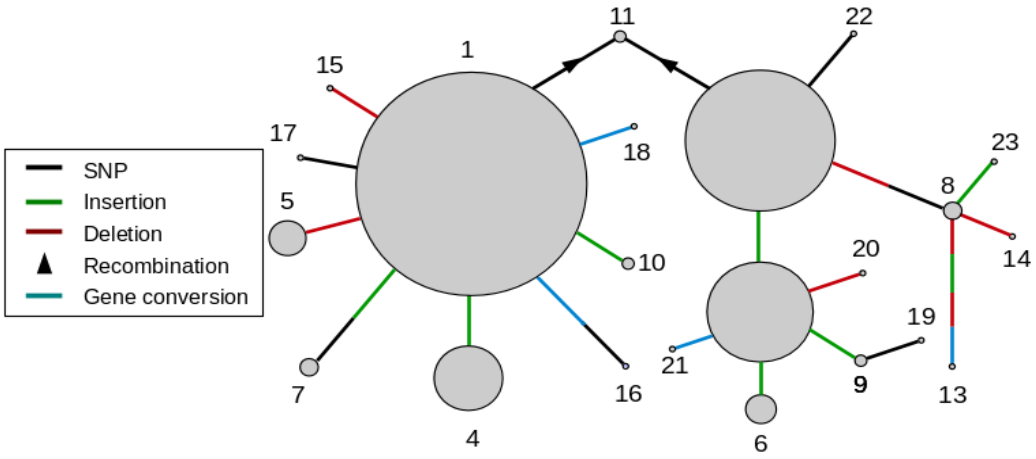

Figure 4: Reconstructed recent history of the GRR

A network showing the inferred relationship between different GRR haplotypes (circles), with circle area corresponding to the frequency in our sample of the DGRP, and connectors representing different mutation or recombination events. Note that there are some haplotypes whose relationship is not easily linked with the others, for example, GRR Hap13 is unlike any other haplotype sequence and there are large differences between GRR Hap1 and GRR Hap2. In other cases it is not clear whether convergent mutation or recombination produced a particular haplotype - for example, the different GRR1 variants each occur in the background of multiple GRR2 variants (see Figure 2).

We also analysed 8 Sanger-sequenced GRR2 haplotypes from a Kenyan population of *D. melanogaster* (Obbard *et al.*, 2006) and compared them to the DGRP haplotypes (Figure 3). Notably, 7 of the 8 Kenyan haplotypes were distinct from each other, and in these 7 haplotypes, four new repeat units were found (GRR2-M, GRR2-N, GRR2-O, GRR2-P; Figure 3). This may suggest that the diversity in the DGRP is a subset of African diversity, as expected from the evolutionary history of this species (Lachaise and Silvain, 2004; Stephan and Li, 2007). GRR2-L, the defining repeat unit of the GRR alpha clade found in the DGRP (gold branches in Figure 3) was rare in the sample of 8 Kenyan sequences, although not absent, suggesting the presence of substantial population structure in GRR.

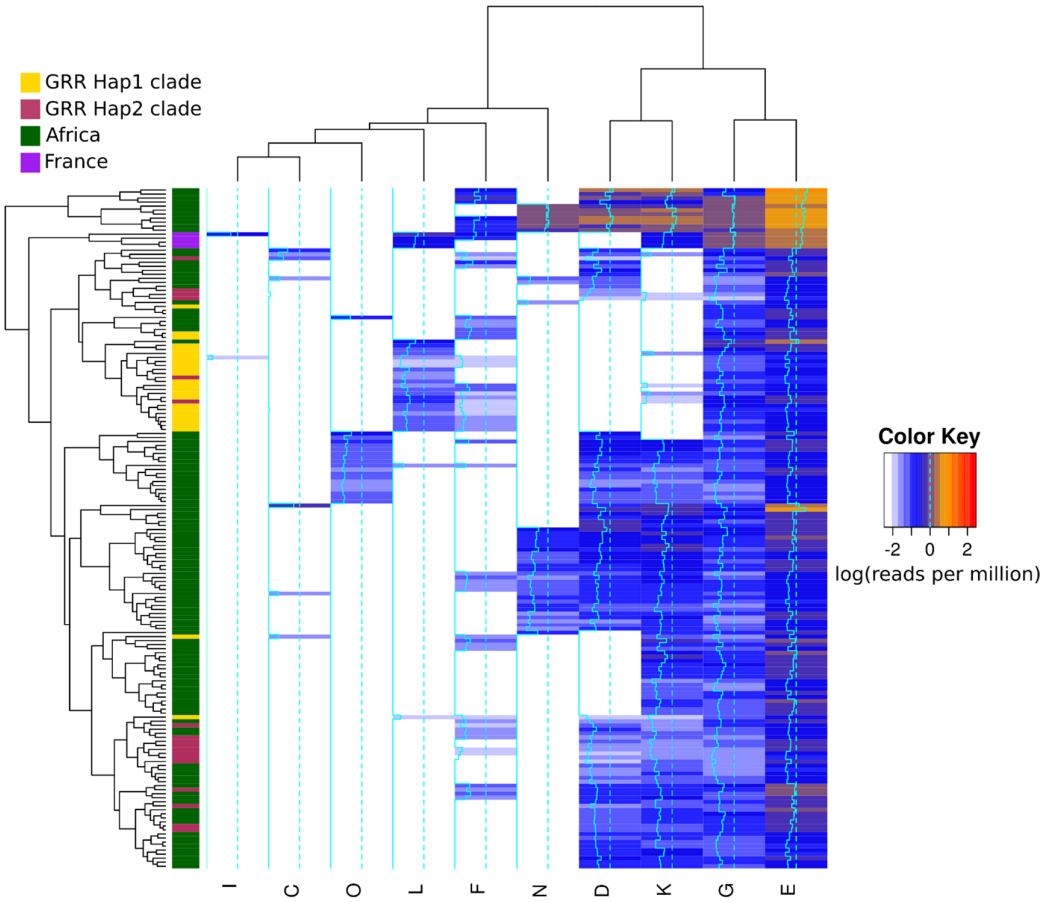

Figure 5: Repeat units in the Drosophila Nexus lines

Clustering of the distribution of repeat units in short read data for a sample in Drosophila populations taken from the nexus dataset (Lack et al, 2015). Lines were excluded if no short reads were found for ubiquitous repeat units (see materials and methods). GRR alpha clade and GRR beta clade are those found in the clustering analysis of Figure 2. DGRP GRR alpha clade appears to be derived from an ancestral African population, whereas GRR beta clade is more divergent, and represents a subset of African diversity. Also, notice the existence of population-specific repeat units (e.g. repeat units O and N) and population specific co-occurrence of repeats (e.g. repeat units L and K in France).

Although we were unable to reconstruct reliable GRR haplotypes from short-read data, we were able to identify the presence of specific repeat units such as *GRR-L*, which in the DGRP is diagnostic of the *GRR alpha* clade, and GRR-D and K, which are diagnostic of *GRR beta* clade. We therefore took advantage of the recent release of the *Drosophila* Genome Nexus, which includes the DGRP as well as individuals sequenced from Africa and France, (Lack *et al.*, 2016) and characterised the distribution of repeat units in these lines (Figure 5). There are repeat units specific to both African (*GRR2-O* and *GRR2-N*) and North American

(*GRR2-B, GRR2-H, GRR2-I,* and *GRR2-J*) populations, although those peculiar to North America were all rare variants. However, French lines also cluster together, characterised by co-occurrence of *GRR2-L* and *GRR2-K* – the defining features of each of the two large classes defined by the DGRP, indicating these lines may be recombinants or heterozygotes. Short read data also suggested that *GRR2-L* is rare in Africa, whereas *GRR-D/K* are common and often co-occur. These observations indicate that the *GRR alpha* clade has risen in frequency since *D. melanogaster* arrived in N. America.

2.4.3 Associations between GRR haplotypes and survival during virus infection

The role of the GRR region is unknown, but as Ago2 is a major effector of antiviral immunity in *Drosophila* (Van Rij *et al.*, 2006), it could function during antiviral defence. Using previously published survival data, we found no significant association between GRR haplotype and resistance to Flock House Virus (95% CI for GRR effect: [-0.12, 0.4], *MCMCp* = 0.30) or Sigma Virus (95% CI for GRR effect: [-0.15, 0.03], *MCMCp* = 0.242) infection in the DGRP. However, when fitting GRR haplotype as a fixed effect, we found that *Hap3* alleles increased longevity following challenge with Drosophila C Virus (DCV) by approximately 0.7 days relative to *Hap1* alleles (*MCMCp = 0.012*, [95% CI 0.21 to 1.17] days*;* Figure S7). This appears to be due to the GRR2 region, as inclusion or exclusion of GRR1 state had no effect. A second model in which GRR1 data were excluded, identified both *Hap2* and *Hap3* as significantly increasing survival relative to Hap1 (*Hap2: MCMCp* = 0.006, 0.56 [0.15 to 0.97] days); Hap3: *MCMCp* = 0.006, 0.64 [0.23 to1.07] days). However, the observed effect is small relative to the effect of the known resistance variant pastrel^{T} (Magwire *et al.*, 2012), which increases longevity in the same experiment by 2.07 days (MCMCp <0.001, [1.58 to 2.54]).

Given the small size of the effect, multiple tests across viruses, and marginal p-values, we elected to perform a second independent test using a subset of the recombinant inbred lines provided by the Drosophila Synthetic Population Resource (DSPR) (King, Macdonald and Long, 2012). An equal number of DSPR lines with haplotype group alpha or haplotype group beta were infected with DCV, and mortality was monitored over a week to determine whether GRR genotype was associated with differences in DCV-induced mortality. In this experiment, although mean survival time was slightly greater for haplotype group beta than group alpha, this trend was not significant (pMCMC = 0.646; Figure S7). The same was true if parent of origin was used as a fixed effect instead of GRR genotype. This is unlikely to be due simply to low power, as we were able to detect a significant association with genotype at the

(albeit larger effect) resistance locus *pastrel* (pMCMC < 0.001). We are therefore unable to replicate the nominally significant effect of GRR haplotype on survival in the DGRP.

2.5 Discussion

2.5.1 GRR amino acid composition is conserved, but repeat unit sequence is not

We observe a high degree of sequence divergence in the Ago2 GRR across insect species. Even over very short timescales, there is high divergence in copy number and repeat unit sequence (Figure 1). This could be explained by a high rate of partial inter-repeat replication slippage, causing the creation of new repeat units from the existing ones, and making the sequence unrecognisable in a relatively short period of time (e.g. Dmel and Dsim GRR2 sequences, which are highly divergent despite only 2.5 My since they shared a common ancestor). In contrast to the sequence of the GRR, we find that the amino acid composition is conserved across the insect species analysed. Although mutational processes can drive sequence conservation in microsatellites, the required concurrent conservation of reading frame in the GRR makes this possibility unlikely. Based on these observations, we envision a scenario where stabilising selection acts at the level of amino acids (e.g. to maintain a certain charge or hydrophobicity) but either lack of constraint or positive selection acts at the level of repeat unit sequence and structure.

2.5.2 GRR repeats are highly variable within *D. melanogaster*, and may be under selection

Repeat number polymorphism in the Ago2 GRR of laboratory lines was previously reported by Hain *et al.* (2010), and our long-read sequencing of a natural population of *D. melanogaster* confirms that this variation is also widespread in the wild. However, our sequencing demonstrates considerable length convergence among haplotypes, such that only 7 different haplotype lengths were present among the 23 distinct haplotypes, and 8 of the 23 distinct haplotypes had the same length (1.035 kbp; Figure 3)—including haplotypes in both the *alpha* and *beta* GRR groups. We found that the haplotypes falling into the *alpha* clade appear to have recently increased in frequency in the North American (DGRP) population. This is supported by a lower diversity surrounding *GRR alpha* than *GRR beta* clade haplotypes, despite the expectation that neutral diversity in linked regions should scale positively with the frequency of the allele. The increase in frequency of the *GRR alpha* clade could be due to drift (e.g. during a bottleneck) or selection, such as parasite-mediated selection acting on

Ago2 GRR repeat region itself. However, given the known selective history of *Ago2* (Obbard *et al.*, 2011), this distribution of haplotype frequencies could also be explained by incomplete linkage to a nearby hard sweep carrying GRR Hap1 to a high frequency (Schrider *et al.*, 2015).

2.5.3 Ago2 GRR variation is not strongly associated with survival after viral challenge

In other genes, extended low-complexity tracks of a single amino acids have known functions, including having been implicated in transcription factor binding (e.g. glutamine, proline, alanine), protein aggregation (glutamine), and cellular localization (histadine), and recently the Q-rich opa repeats of *Notch* have been found to be involved in developmental defects (Gerber *et al.*, 1994; Salichs *et al.*, 2009; Gemayel *et al.*, 2015; Rice *et al.*, 2015). But, although the long-term conservation of the Ago2 GRR among pancrustacea argues that it is maintained by selection, the function of this repeat region remains unclear. As viral suppressors of RNAi (VSRs) have been proposed as the likely drivers of the rapid protein evolution of Ago2 (Obbard et al., 2009), and high diversity is predicted by many models of host-parasite coevolution (Anderson and May, 1982; Antonovics and Thrall, 1994; Sasaki, 2000; Ebert, 2008; Brockhurst *et al.*, 2014) it is tempting to speculate that the Ago2 GRR may play a role in VSR evasion. For example, the GRR could act to cover residues that underlie Ago2-VSR interactions, or as a bait region, sequestering VSRs away from the catalytic residues of Ago2. However, although *Ago2* GRR showed a significant association with survival after *DCV* infection in our re-analysis of published data from 127 of the DGRP lines, we were unable to replicate this using selected lines from the DSPR. These conflicting results may reflect a false positive from the DGRP analysis, or low power in the DSPR analysis, perhaps due to the challenge inherent in categorising GRR haplotypes using short read data. However, in either case, it is clear any association must be weak relative to previously identified segregating functional polymorphisms, such as *pastrel*.

2.5.4 The potential importance of complex repeat sequences in GWAS studies

We find that LD within the GRR, and between the GRR and surrounding variants, is low (Figure S4), indicating that any phenotypic association with this repeat region would be difficult to identify through GWAS using linked sites only. Additionally, the convergence in length between highly divergent GRR haplotypes means that simple length assays may not be suitable to differentiate between haplotypes and may miss important variants. More generally, our study suggests that short read sequencing, such as that currently employed by the majority

of association studies, is not a viable option for repetitive regions, as we were only able to assemble one correct *Ago2* GRR haplotype among the 117 DGRP datasets using public sequence read data. Clustering by repeat unit presence in short read data confirm our PacBio-sequenced haplotypes (Figure 5), but may only be useful if there is prior knowledge to the possible repeat units in a population and if the region is sequenced in high depth. For example, reads with repeat units *GRR2*-A, *GRR2-G* and *GRR2-E* (which occur in every haplotype) were not always detectable in the short read data for a sample. This indicates that GRR coverage can be low, that incorrect haplotype inference was not only due to assembly errors, and may indicate that the GRR region has unusually low coverage – perhaps because it is not conducive to short read sequencing. Together, these attributes argue that sequencing repetitive regions can provide a depth of understanding not attainable by looking at length variation alone.

Chapter 3: Adaptive evolution of RNAi genes in invertebrates

The text of this chapter is published:

Palmer, William H., Jarrod D. Hadfield, and Darren J. Obbard. "RNA-interference pathways display high rates of adaptive protein evolution in multiple invertebrates." Genetics 208.4 (2018): 1585-1599.

I wrote this chapter with comments and text edits from Darren Obbard and Jarrod Hadfield. In addition to the supplementary figures and data provided in an appendix, supplemental R code, tables, and further model descriptions are available through Figshare (DOI: 10.6084/m9.figshare.5843991.v1) and in the supplementary material of Palmer et al, (2018).

3.1 Abstract

Conflict between organisms can lead to reciprocal adaptation that manifests itself as an increased evolutionary rate in genes mediating the conflict. This adaptive signature has been observed in RNA interference (RNAi) pathway genes involved in the suppression of viruses and transposable elements in *Drosophila melanogaster*, suggesting that a subset of *Drosophila* RNAi genes may be locked into an arms race with these parasites. However, it is not known whether rapid evolution of RNAi genes is a general phenomenon across invertebrates, or which RNAi genes generally evolve adaptively. Here we use population genomic data from eight invertebrate species to infer rates of adaptive sequence evolution, and to test for past and ongoing selective sweeps in RNAi genes. We assess rates of adaptive protein evolution across species using a formal meta-analytic framework to combine data across species, and by implementing a multispecies generalised linear mixed model of mutation counts. Across species, we find that RNAi genes display a greater rate of adaptive protein substitution than other genes, and that this is primarily mediated by positive selection acting on the genes most likely to defend against viruses and transposable elements. In contrast, evidence for recent selective sweeps is broadly spread across functional classes of RNAi genes, and differs substantially among species. Finally, we identify genes that exhibit elevated adaptive evolution across the analysed insect species, perhaps due to concurrent parasite-mediated arms races.

3.2 Introduction

RNA-interference mechanisms include a diverse group of pathways, united by their use of Argonaute-family proteins complexed with short (20-30 nt) RNA molecules to target longer RNA molecules through sequence complementarity (Carmell *et al.*, 2002; Meister, 2013). These pathways regulate multiple biological processes that can be divided into three distinct subpathways in arthropods and nematodes, each represented by a characteristic class of small RNAs: the micro-RNA (miRNA), the short-interfering RNA (siRNA), and the piwi-interacting RNA (piRNA) pathways. The miRNA pathway processes endogenously-encoded fold-back hairpins which, in their mature miRNA form, regulate gene expression and coordinate developmental processes (Alvarez-Garcia and Miska, 2005; Chen *et al.*, 2014; Ha and Kim, 2014). The siRNA pathway has two distinct roles, depending on the endogenous or exogenous origin of its substrate. First, the endo-siRNA pathway processes endogenously encoded dsRNA to regulate processes such as TE defense (Czech *et al.*, 2008; Ghildiyal *et al.*, 2008; Kawamura *et al.*, 2008) chromosomal segregation (Hall, Noma and Grewal, 2003; Huang *et al.*, 2015), and heterochromatin formation (Deshpande, Calhoun and Schedl, 2005). Second, the exo-siRNA (or viRNA) functions primarily as a form of antiviral immunity (Wang *et al.*, 2006; Bronkhorst and van Rij, 2014). The piRNA pathway forms a defence against transposable elements (TEs), which is germline limited in some species, and is mediated by piRNAs derived from endogenously-encoded clusters of inactivated TE sequences and from active TEs (Klattenhoff and Theurkauf, 2008; Thomson and Lin, 2009; Czech *et al.*, 2016; Lewis *et al.*, 2018).

Within this framework there is substantial mechanistic variation among species, and RNAi-pathway components seem to be evolutionarily labile. For example, in nematodes the mechanism and function of the piRNA pathway is not well conserved: primary piRNA-like small RNAs are encoded by short distinct loci instead of the clusters observed in flies and mammals, and mediate the biogenesis of a separate endo-siRNA population transcribed by an RNA-dependent RNA Polymerase (RdRP) and processed by Dicer (Duchaine *et al.*, 2006; Das *et al.*, 2008). Further, only one of the five major clades of nematode have retained Piwi-subfamily proteins — the canonical effector of the piRNA pathway —and instead rely solely on the (RDRP-produced) endo-siRNAs (Sarkies *et al.*, 2015). The piRNA pathway can also take on entirely new roles, for example, duplications of *piwi* in *Aedes* mosquitoes has allowed the piRNA pathway to adopt an antiviral role in somatic tissues (Morazzani *et al.*, 2012), while

other *piwi* duplicates maintain the ancestral function (Miesen, Girardi and van Rij, 2015; Miesen *et al.*, 2016).

The role of RNAi pathways in mediating inter-genomic (host-virus) and intra-genomic (host-TE, segregation distortion) (Ferree and Barbash, 2007) conflict may make them a target of antagonistic host-parasite coevolution. This could result in balancing or directional selection on the loci in conflict, evidenced by the maintenance of polymorphism or elevated rates of adaptive fixation, respectively (Anderson and May, 1982; Ebert, 2008). This has been well studied in *Drosophila* RNAi pathway genes, which show elevated rates of adaptive protein evolution (Obbard *et al.*, 2006; Obbard *et al.*, 2009), signatures of selective sweeps(Kolaczkowski, Hupalo and Kern, 2011; Obbard *et al.*, 2011; Lewis *et al.*, 2016), and sites with elevated protein evolution across the *Drosophila* phylogeny (Vermaak, Henikoff and Malik, 2005; Heger and Ponting, 2007; Kolaczkowski, Hupalo and Kern, 2011). For example, a comparison of the antiviral RNAi genes *AGO2*, *Dcr-2*, and *r2d2* to their miRNA functional counterparts with no known role in conflict (the paralogs *AGO1*, *Dcr-1*, and *loqs*) shows a striking difference in rates of protein evolution, as well as a greater rate of adaptive amino-acid substitution (Obbard *et al.*, 2006). In addition, evolutionary rates of piRNA pathway genes involved in transcriptional silencing are elevated and highly correlated with other piRNA pathway genes across the *Drosophila* phylogeny (Blumenstiel, Erwin and Hemmer, 2016).

Although some antiviral and anti-TE RNAi pathway genes clearly display elevated rates of adaptive protein evolution in *Drosophila*, the generality of this pattern remains to be elucidated. Here we apply both traditional McDonald-Kreitman (McDonald and Kreitman, 1991) and SnIPRE-style (Eilertson, Booth and Bustamante, 2012) analyses, and selective sweep-based analyses (Nielsen, et al., 2005; Pavlidis, et al., 2013), to publicly-available genome-scale data from three dipterans, two lepidopterans, a hymenopteran, and two clade V nematodes (Sarkies *et al.*, 2015).

By combining estimates across species, we investigate the specific RNAi subpathways that may be the target of elevated positive selection. This allows us to estimate the rates of adaptation across species, thereby improving single gene estimates and allowing us to identify genes that are undergoing parallel adaptation across the taxa analysed. Finally, we summarise the evidence for recently completed and ongoing selective sweeps in RNAi genes across these eight taxa. We conclude that rapid evolution of RNAi genes is a general phenomenon

in these eight invertebrates, although evidence for recent sweeps is highly contingent on the focal species.

3.3 Materials and Methods

3.3.1 Selection of genes for analysis

Putative RNAi pathway genes of *Drosophila melanogaster* and *Caenorhabditis elegans* were used to find homologues in six insect and two nematode species (Table 1, Table 2). For the six species of insect, we classified these RNAi genes as miRNA, piRNA, siRNA, or viRNA, based on a literature search (Table 1). Therefore, when we refer to insect "RNAi genes", we have used all genes in Table 1 for the analysis, and when we refer to specific subpathways or processes, we have used only those genes designated as in that functional category as per the third column of Table 1. Where a gene was implicated in more than one subpathway, we assigned it to the pathway that has been independently experimentally validated most often. Although the viRNA and siRNA pathways are not easily separable, we make this distinction based on the hypothesis that AGO2, Dcr-2, and R2D2 may be evolving adaptively in response to viruses, as there is direct experimental evidence from independent studies that these genes have an antiviral role in multiple insect species against phylogenetically diverse viruses. We also split the piRNA pathway genes among three functional categories: post-transcriptional silencing effectors, transcriptional silencing effectors, and biogenesis machinery. A gene was considered a biogenesis factor if piRNA levels decrease upon loss-of-function, an effector if piRNA pathway function is compromised without reducing piRNA levels, and a transcriptional silencing effector if the effector is involved in transcriptional silencing (Table 1). Finally, we selected 65 piRNA genes in *D. melanogaster* with known tissue-specificity to calculate rates of adaptation in the germline versus the somatic follicle cells (Table S3). This gene list contains the core of the piRNA pathway and genes independently validated in two of three screens for piRNA pathway constituents (Czech *et al.*, 2013; Handler *et al.*, 2013; Muerdter *et al.*, 2013).

Gene	Flybase ID	Pathway	Example References
AGO1	FBgn0262739	miRNA	Okamura et al, 2004
AGO2	FBgn0087035	viRNA	Wang et al, 2006; van Rij et al, 2006
AGO3	FBgn0250816	piRNA - effector	Brennecke et al, 2007; Gunawardane et al, 2007
armi	FBgn0041164	piRNA - biogenesis	Saito et al, 2010
Ars2	FBgn0033062	miRNA	Gruber et al, 2009; Sabin et al 2009
arx	FBgn0036826	piRNA - effector	Ohtani et al, 2013; Donertas et al, 2013
bel	FBgn0263231	piRNA	Lo et al, 2016
csul	FBgn0015925	piRNA - effector	Kirino et al, 2009
cuff	FBgn0260932	piRNA - transcriptional	Pane et al, 2012; Mohn et al, 2014
Dcr-1	FBgn0039016	miRNA	Jiang et al, 2005; Saito et al, 2005
Dcr-2	FBgn0034246	viRNA	Wang et al, 2006; Galiana-Arnoux et al, 2006
del	FBgn0086251	piRNA - transcriptional	Mohn et al, 2014; Parhad et al, 2017
drosha	FBgn0026722	miRNA	Lee et al, 2003
egg	FBgn0086908	piRNA - transcriptional	Rangan et al, 2011; Sienski et al, 2015
FMR1	FBgn0086908	siRNA	Ishizuka et al, 2002
Hel25E	FBgn0014189	piRNA - biogenesis	Zhang et al, 2012; Zhang et al 2014
hen1	FBgn0033686	piRNA - biogenesis	Saito et al, 2007; Horwich et al, 2007
krimp	FBgn0034098	piRNA	Sato et al, 2015
loqs	FBgn0032515	miRNA	Jiang et al, 2005; Saito et al, 2005
mael	FBgn0016034	piRNA - transcriptional	Sienski et al, 2012
mei-p26	FBgn0026206	miRNA	Neumuller et al, 2008
papi	FBgn0031401	piRNA - biogenesis	Saxe et al, 2013; Izumi et al, 2016
pasha	FBgn0039861	miRNA	Yeom et al, 2006
piwi	FBgn0004872 FBgn0000146	piRNA - effector	Brennecke et al, 2007
qin	FBgn0263974	piRNA - biogenesis	Zhang et al, 2011
r2d2	FBgn0031951	viRNA	Liu et al, 2003
rhi	FBgn0004400	piRNA - transcriptional	Klatenhoff et al, 2009; Mohn et al, 2014
rm62	FBgn0003261	siRNA	Csink et al, 1994; Boeke et al, 2011
shu	FBgn0003401	piRNA - biogenesis	Olivieri et al, 2012; Preall et al, 2012
spn-E	FBgn0003483	piRNA - biogenesis	Malone et al, 2009; Ryazansky et al, 2016
squ	FBgn0267347	piRNA - effector	Haase et al, 2010
tapas	FBgn0027529	piRNA - biogenesis	Patil et al, 2014
tejas	FBgn0033921	piRNA - biogenesis	Patil et al, 2010
trax	FBgn0038327	siRNA	Liu et al, 2009
trsn	FBgn0033528	siRNA	Liu et al, 2009
TSN	FBgn0035121	siRNA	Caudy et al, 2003
tudor	FBgn0003891	piRNA - biogenesis	Nishida et al, 2009
vasa	FBgn0283442	piRNA - biogenesis	Xiol et al, 2014
vig	FBgn0024183	siRNA	Caudy et al, 2002
vret	FBgn0263143	piRNA - biogenesis	Zamparini et al, 2011
wde	FBgn0027499	piRNA - transcriptional	Koch et al, 2009
Yb	FBgn0000928 FBgn0051755 FBgn0037205	piRNA - biogenesis	Saito et al, 2010
zuc	FBgn0261266	piRNA - biogenesis	Mohn et al, 2015; Han et al, 2015

Table S1: Insect RNAi genes

A table with the gene names, FlyBase identifiers, subpathway involvement, references, and motivation for sub-pathway involvement for each insect gene analysed. See Table S1 in Palmer et al (2018, DOI: 10.1534/genetics.117.300567) for a more detailed description.

Gene	C. elegans Wormbase ID	Gene	C. elegans Wormbase ID	Gene	C. elegans Wormbase ID
ACR-11	WBGene00000050	**NHL-2**	WBGene00003598	**WAGO-5**	WBGene00022877
AIN-2	WBGene00015007	**NRDE-1**	WBGene00007577	**Y23H5A.3**	WBGene00021270
C04F12.1	WBGene00007297	**RDE-4**	WBGene00004326	**ERGO-1**	WBGene00019971
CDC-25.1	WBGene00000386	**SID-2**	WBGene00004796	**NRDE-2**	WBGene00011333
DRH-3	WBGene00008400	**ZK418.8**	WBGene00022737	**PHO-1**	WBGene00004020
DRSH-1	WBGene00009163	**CSR-1**	WBGene00017641	**PID-1**	WBGene00017549
ERI-6	WBGene00016561	**DRH-1**	WBGene00001090	**PIR-1**	WBGene00011967
ERI-7	WBGene00016566	**EPI-1**	WBGene00001328	**RRF-3**	WBGene00004510
F57C9.7	WBGene00019013	**ERI-1**	WBGene00001332	**VIG-1**	WBGene00006924
GEI-11	WBGene00001568	**ERI-5**	WBGene00021419	**WAGO-11**	WBGene00021711
HAF-6	WBGene00001816	**LAM-1**	WBGene00002247	**ALG-3**	WBGene00011910
HPO-24	WBGene00011945	**RSD-2**	WBGene00004681	**ALG-4**	WBGene00006449
MEL-26	WBGene00003209	**F20A1.9**	WBGene00017620	**CGH-1**	WBGene00000479
MUT-16	WBGene00003508	**MUT-15**	WBGene00011323	**MUT-7**	WBGene00003504
MUT-2	WBGene00003499	**RDE-1**	WBGene00004323	**TBP-1**	WBGene00006542
PTR-2	WBGene00004217	**SID-1**	WBGene00004795	**Y37D8A.1(**	WBGene00012554
RDE-10	WBGene00021634	**WAGO-4**	WBGene00010263	**COGC-2**	WBGene00000585
RRF-1	WBGene00004508	**LAM-2**	WBGene00016913	**ERI-3**	WBGene00021103
RRF-2	WBGene00004509	**C35E7.8**	WBGene00016460	**RDE-11**	WBGene00023421
RSD-6	WBGene00004684	**EGO-1**	WBGene00001214	**C14B1.7**	WBGene00007578
ULP-5	WBGene00006740	**EKL-1**	WBGene00009052	**C27H6.3**	WBGene00007785
W01A8.5	WBGene00012167	**PPW-1**	WBGene00004093	**DCS-1**	WBGene00000940
PASH-1	WBGene00011908	**PPW-2**	WBGene00004094	**HRDE-1**	WBGene00007624
MEL-47	WBGene00017132	**PRG-1**	WBGene00004178	**MUT-14**	WBGene00003507
TSN-1	WBGene00006626	**PRG-2**	WBGene00004179	**NRDE-3**	WBGene00019862
XRN-2	WBGene00006964	**RDE-2**	WBGene00004324	**PRDE-1**	WBGene00008995
CUL-2	WBGene00000837	**SAGO-1**	WBGene00019666	**RDE-12**	WBGene00010280
DCR-1	WBGene00000939	**SAGO-2**	WBGene00018921	**WAGO-10**	WBGene00020707
ERI-9	WBGene00016143	**SMG-2**	WBGene00004880	**AIN-1**	WBGene00015547
HENN-1	WBGene00015349	**WAGO-1**	WBGene00011061	**ALG-1**	WBGene00000105
HPL-2	WBGene00001996	**WAGO-2**	WBGene00018862	**ALG-2**	WBGene00000106

Table S2: Nematode RNAi genes

A table with gene names and WormBase identifiers for the genes used in C. briggsae and P. pacificus.

Homologs of the *D. melanogaster* and *C. elegans* genes were identified using a two-step process. First, a hidden Markov Model (HMMer) (Eddy, 2008) was used to find best reciprocal best-hits for a gene of interest using predicted protein sets (if available) or UniProtKB. If no hit was found, then Exonerate was used to identify unannotated homologues in the genome using the model 'protein2genome' (Slater and Birney, 2005). If exonerate was unable to model a homologue, then this gene was classified as missing, either due to gene loss or an incomplete genome assembly. We defined genes as duplicates (paralogues) if multiple regions of a genome shared a best hit to a reference gene, and these regions showed substantial sequence divergence between them (i.e. they were not obviously a mis-assembly duplicate or allelic). Because the large divergence times between insects and nematodes and the complexity of RNAi pathways in nematodes make homology assignment uncertain, we restricted our gene-level analyses to insects.

3.3.2 Population genomic data

We used previously published population genomic data for *Drosophila melanogaster* (Lack *et al.*, 2016)*, Drosophila pseudoobscura* (Pseudobase) (McGaugh *et al.*, 2012)*, Anopheles gambiae* (The *Anopheles gambiae* 1000 Genomes Consortium, 2014: Ag1000G phase 1 AR2 data release. MalariaGEN.)*, Heliconius melpomene* (Kronforst *et al.*, 2013)*, Bombyx mandarina* (Xia Qingyou, 2009)*, Apis mellifera* (Harpur *et al.*, 2014)*, Pristionchus pacificus* (Rödelsperger *et al.*, 2014)*,* and *Caenorhabditis briggsae* (Thomas *et al.*, 2015) for our analyses (Table S4). For both *Drosophila* species, we used previously-published haplotype data (haploid sequencing of *D. melanogaster*, inbred lines of *D. pseudoobscura*). For the other taxa we obtained raw sequencing reads from EBI ENA (identifiers provided in Table S4) and mapped them to the most recent reference genome for each species using Bowtie2 (Langmead and Salzberg, 2012) with default settings. We used GATK's HaplotypeCaller on each individual separately (DePristo *et al.*, 2011) to call variants in a 200 kb region surrounding each gene of interest (i.e. 100 kb either side of the RNAi gene, unless the contig was less than 200 kb). For high coverage datasets (*A. mellifera, H. melpomene, C. briggsae, A. gambiae,* and *P. pacificus*) we excluded sites with a read depth lower than 5, but we reduced this threshold to 2 for the low-coverage *B. mandarina*. All sites above this filter were included in the analysis. After mapping and filtering sites we created two randomly resolved pseudohaplotype sequences per individual (i.e. without any phase information) from the sites that remained, and these were used for downstream analyses (none of which depend on phase). Only one haplotype was sampled from each *C. briggsae* and *P. pacificus* individual, as the sequenced individuals were reported to be highly homozygous. In *H. melpomene*, we occasionally observed long stretches of high divergence shared by multiple individuals. We assumed these to be possible cases of either contamination, inversions that have recently risen to a high frequency, or introgression (Pardo-Diaz *et al.*, 2012), and removed these haplotypes.

To calculate divergence between genes, and to polarise mutations for sweep analyses, we used the outgroup species *Drosophila simulans, Drosophila miranda, Heliconius hecale, Bombyx huttoni, Anopheles christyi* and *Anopheles melas, Apis cerana, Caenorhabditis nigoni,* and *Pristionchus exspectatus*, respectively (Table S4). Outgroups were selected based on their divergence from the ingroup species (*ca.* 1-10% divergence of all sites) and on the availability of genomic data. For *A. gambiae* we tested outgroups with low (*An. melas*) and high

(*An. christyi*) divergence times, as most *Anopheles* species are too close or too divergent to provide a robust outgroup for MK tests (Obbard *et al.*, 2007), and our results remain qualitatively the same for both outgroups (*A. melas* used for the presented analyses). For *D. simulans* (FlyBase, r2.02)*, D. miranda* (Pseudobase, MSH22 strain)*, A. melas* (VectorBase, CM1001059 strain, AmelC1 assembly)*, A. christyi* (VectorBase, ACHKN1017 strain, AchrA1 assembly)*, B. huttoni* (Sackton *et al.*, 2014) (BioProject PRJNA198873), and *P. exspectatus* (WormBase, Bioproject PRJEB6009), the outgroup reference assemblies were publicly available and used as provided. However, the *Caenorhabditis nigoni* reference assembly sequence is contaminated with the more divergent nematode *Caenorhabditis afra* (Thomas *et al.*, 2015), and *Caenorhabditis nigoni* is the only current suitable outgroup for C. briggsae. We therefore applied a sliding window across the alignments between *C. nigoni* and *C. afra*, and arbitrarily excluded regions that were greater than 6 standard deviations from the mean divergence. Published reference assemblies were not available for *Apis cerana* and *Heliconius hecale*. To generate outgroup sequences for these species we iteratively remapped reads (*H. hecale:* ERR260306; *A. cerana:* SRR957079) to the respective *Apis mellifera* and *Heliconius melpomene* references, each time updating the previous reference with homozygous non-reference calls. These reads were mapped with Bowtie2 and then remapped with the divergent alignment software, Stampy (Lunter and Goodson, 2011). Homozygous nonreference calls (enriched for sites divergent between the ingroup and outgroup) were made with GATK's HaplotypeCaller, with the heterozygosity parameter set to the expected divergence between species. Such sequences will not perfectly reflect the true outgroup sequence, and are expected to be biased toward the ingroup, downwardly biasing estimates of divergence in high-divergence regions. However, we confirmed that this approach works well by iteratively mapping *D. simulans* to *D. melanogaster*, and comparing the result with the known *D. simulans* assemblies (K_S= 0.10 for iterative mapping vs K_S=0.12 for the true assembly), and while bias probably remains, it is unlikely to spuriously elevate the inferred rates of one class of genes relative to the other. More generally, our approach to mapping, filtering, and variant calling may be prone to such biases, but they are unlikely to differentially affect gene classes of different function.

For the MK-based analyses, target sequences were aligned as amino acids using MUSCLE (Edgar, 2004), and then each gene alignment examined manually to remove putative misalignments. Likely misalignments were identified by eye as regions of unusually high divergence with no amino acid similarity to the consensus sequence, often occurring at the ends

of the gene. We assumed these were caused by miss-assembly, and removed these blocks from the alignment. Within-species data was aligned first, and then a consensus sequence of this alignment used to align against the outgroup sequence. Synonymous and nonsynonymous substitutions between species were inferred using codeml from the PAML package using the YN00 model (Yang and Nielsen, 2000), which estimates substitution rates using an approximation to maximum likelihood methods, while accounting for base composition differences between codon positions and differences in transition/transversion rates.

3.3.3 Rates of adaptive protein evolution by pathway

To estimate the rate of adaptive protein evolution in different functional classes of gene, and to test for differences in rate between classes, we used two different approaches derived from the McDonald-Kreitman test ('MK framework') (McDonald and Kreitman, 1991). The MK framework combines polymorphism and divergence data from putatively unconstrained (synonymous) and potentially selected (nonsynonymous) variants to infer an excess of nonsynonymous fixations that can be attributed to positive selection. This framework was later formalised in several maximum likelihood and Bayesian methods to estimate α, the proportion of nonsynonymous substitutions that are adaptive (Charlesworth, 1994; Fay, Wyckoff and Wu, 2001, 2002; Smith and Eyre-Walker, 2002; Sawyer *et al.*, 2003; Bierne and Eyre-Walker, 2004; Welch, 2006). However, α and related statistics can be biased by slightly deleterious mutations. This is because such mutations are unlikely to fix, but do contribute substantially to polymorphism (McDonald and Kreitman, 1991; Eyre-Walker, 2002; Charlesworth and Eyre-Walker, 2008; Eyre-Walker and Keightley, 2009; Gossmann, Keightley and Eyre-Walker, 2012). We used DFE-alpha and SnIPRE to estimate rates of adaptive evolution. These complementary approaches model the population genetic processes responsible for these biases (DFE-alpha) or the resulting genome-wide variability caused by these biases (SnIPRE) (Eyre-Walker and Keightley, 2009; Eilertson, Booth and Bustamante, 2012).

In the first approach, we used an explicit population-genetic model to estimate the number of adaptive nonsynonymous substitutions per site (DFE-alpha) (Eyre-Walker and Keightley, 2009). This approach has the advantage that it provides direct estimates of the parameters of interest, and explicitly models changes in population size (as reflected by the site frequency spectrum of unconstrained sites) and the distribution of deleterious fitness effects, which might otherwise bias estimates (Keightley and Eyre-Walker, 2007; Eyre-Walker and Keightley, 2009). However, as currently implemented, this method does not allow data to be

directly combined among species. Therefore, to obtain more precise homologue- and pathway-based estimates we combined per-gene point estimates from DFE-alpha using a linear mixed model (including their estimated uncertainty; i.e. a meta-analysis). In the second approach, we used an extension of the SnIPRE model (Eilertson, Booth and Bustamante, 2012), which re-frames the MK framework as a linear model in which polymorphism and substitution counts are predicted by synonymous or nonsynonymous state. Although this model does not explicitly consider the same underlying population-genetic processes, it does permit a straightforward extension to natively include gene, homologue, pathway, and host species as predictors, and therefore provides a direct test of the questions of interest (although at a cost of potentially less accurate or arbitrarily-scaled parameter estimates; see text S1). We have re-implemented the SnIPRE model using the Bayesian Generalised Linear Mixed Modelling R package MCMCglmm (Hadfield, 2010).

DFE-alpha analyses

DFE-alpha (Eyre-Walker and Keightley, 2009). infers ω_A (the number of adaptive nonsynonymous substitutions per nonsynonymous site, relative to the number of synonymous substitutions per synonymous site), while simultaneously modelling the distribution of deleterious fitness effects and a population size change (Keightley and Eyre-Walker, 2007; Eyre-Walker and Keightley, 2009). The ω_A statistic is closely related to the more widely reported α statistic (Charlesworth, 1994; Fay, Wyckoff and Wu, 2001, 2002; Smith and Eyre-Walker, 2002; Sawyer *et al.*, 2003; Bierne and Eyre-Walker, 2004; Welch, 2006), but differs in that ω_A is expected to be less dependent on effective population size and therefore better for cross-species comparisons. This is because the denominator, dS, should be less affected by the efficacy of selection, and thus effective population size (Gossmann *et al.*, 2010; Gossmann, Keightley and Eyre-Walker, 2012; Kousathanas, Halligan and Keightley, 2014). DFE-alpha utilises the observed site frequency spectrum (SFS) for putatively unconstrained synonymous sites and potentially selected nonsynonymous sites, and maximises the likelihood of observing these spectra given the distribution of deleterious fitness effects (DFE) for nonsynonymous variants and a step-change in effective population size (Eyre-Walker and Keightley, 2009). The 'excess' nonsynonymous divergence attributable to adaptive substitution is then inferred, given the maximum likelihood estimate of the DFE and the observed divergence (Eyre-Walker and Keightley, 2009). We inferred ω_A for: (i) each RNAi gene and each position-matched 'control' gene (i.e. those with no known RNAi-pathway role falling within the same

200 Kbp interval); (ii) each RNAi subpathway and their respective control genes, and; (iii) all RNAi pathway genes together, by pooling polymorphism and divergence data across genes within classes. We then compared this grouped polymorphism and divergence data in pathways of interest against control genes. We estimated the parameters of the nominal change in population size (the relative population size change parameter N_2, and the time of the population size change, t_2) for all genes treated together within species, and then fixed these estimates for pathway and individual gene estimates. Conditional on this species-wide estimate of demographic history, the DFE was estimated separately for RNAi and control genes. We obtained confidence intervals for estimates of α and ω_A by bootstrapping genes within classes (1000 draws), and we tested for differences in rate between gene classes by randomly permuting genes 1000 times between classes. To test for differences in the DFE between RNAi and control genes we performed a likelihood ratio test between a null model in which parameters of the DFE were estimated for all genes together, and an alternative one in which we allowed the DFE parameters to be estimated separately for RNAi and control genes, holding all other parameters constant.

Pooling polymorphism and divergence data across genes allows calculation of pathway-specific ω_A within a species, but cannot readily give cross-species estimates. Further, underlying unaccounted-for structure in the data could bias results, for example, if a high divergence gene has few polymorphisms, then it will influence divergence, but not DFE, estimation. Therefore, we also calculated ω_A for individual genes in each species, and analysed these estimates across species. In general, such estimates are extremely poor unless samples sizes are extremely large (e.g. hundreds of alleles are sampled, or genes are very large) (Keightley and Eyre-Walker, 2010). However, if the selective pressure acting on genes is consistent across species, as assumed by many phylogenetic approaches (Yang, 2007), we can acquire more accurate estimates of the relative rate of adaptive evolution by combining information across species. We therefore used a formal meta-analytic approach to combine small-group and single-gene estimates across species by constructing linear mixed models using MCMCglmm (Hadfield, 2010). These models were used to estimate average gene-level ω_A of pathways and homologues, and variation among gene-level ω_A estimates.

The first three models took the same form, only distinguished by the pathways among which genes were divided. In Model 1A the genes were classified as either ‘control’ or ‘RNAi’, in Model 1B the RNAi class was expanded into four levels: ‘miRNA’, ‘siRNA’, ‘viRNA’, and ‘piRNA’

and in Model 1C the piRNA class was further split into three functional categories: 'effectors of transcriptional silencing', 'effectors of post-transcriptional silencing', and 'biogenesis factors'. The model for the estimate of ω_A (i.e. $\widehat{\omega}_A$) for homologue *k* in gene class *l* in species *m* had the form:

$$\widehat{\omega}_{A:klm} = \beta_0 + \beta_{Class:l} + u_{Organism:m} + m_{klm} + \varepsilon_{klm} \qquad [1]$$

where β_0 is the intercept, $\beta_{Class:l}$ is a fixed effect associated with gene class *l*, $u_{Organism:m}$ is a random effect associated with species *m*, m_{klm} is the sampling error associated with each estimate, and ε_{klm} is the between observation error after accounting for measurement error, which was allowed to vary by gene class (i.e. pathway). The variance of the sampling errors was obtained by bootstrapping genes by codon, and this sampling error variance was fixed at that value in the analysis. All species effects were assumed to come from a single normal distribution but the errors were assumed to come from independent normal distributions with different variances for each gene class.

Model 2 extended Model 1 by including homologue as a random effect ($u_{Hom:kl}$) in order to identify homologues with elevated adaptation across lineages, where each homologue *k* is represented by approximately 6 genes, one from each insect species analysed, except for the cases where an orthologue was not present in every species (resulting in fewer than 6 observations for a homologue), or a gene had duplicated in one or more of the species (resulting in greater than 6 observations for a homologue).

$$\widehat{\omega}_{A:klm} = \beta_0 + \beta_{Class:l} + u_{Organism:m} + u_{Hom:kl} + m_{klm} + \varepsilon_{klm} \qquad [2]$$

Here, the homologue effects were assumed to come from independent normal distributions with different variances for each gene class. In this model the cross-species average ω_A for a homologue *k* in gene class *l* is given by $\overline{\omega}_{A:kl} = \beta_0 + u_{Class:l} + u_{Hom:kl}$. However, if genes are misclassified with respect to the gene class they belong, then $\overline{\omega}_{A:kl}$ is likely to biased in general, and particularly so for misclassified genes. An arguably more conservative approach is to only use information from homologous genes to estimate the cross-species (i.e. remove the class effects from the model; this approach is provided as Model 2B in S1 text) and have $\overline{\omega}_{A:kl} = \beta_0 + u_{Hom:kl}$.

SnIPRE-like analysis

The meta-analytic approach to cross-species analysis above has the advantage of utilising DFE-alpha estimates that are inferred under an explicit population-genetic model. However, it has the disadvantage that it conditions on point estimates from a model, rather than using the available data directly. We have therefore taken advantage of the Poisson linear mixed model approach to MK analyses 'SnIPRE' proposed by Eilertson et al. (2012), which models the counts of mutations in four classes: synonymous within-species polymorphisms, nonsynonymous within-species polymorphisms, between-species synonymous differences (divergence) and between-species nonsynonymous differences. By fitting 'nonsynonymous' and 'divergent' as main effects, selection can be inferred from their interaction, which records the excess contribution of nonsynonymous mutations to between-species divergence. This excess can be assessed at the level of individual genes (by treating gene identity as a random effect) or can be expressed as a function of other fixed or random effects such as gene class and species. Although this approach does not directly provide parameter estimates that are interpretable in simple population-genetic terms, such as ω_A, it has the advantage of extending naturally to provide comparisons between species and gene classes while still using raw count data directly. Here we combine polymorphism and divergence data from several species to test whether RNAi genes have higher rates of adaptive substitution than our set of control genes, whether these rates vary between different subclasses of RNAi gene, and whether these rates vary between different homologues. We fitted these models with the R package MCMCglmm (Hadfield, 2010) and the code is provided in the S1 text. In their single-species and single-class analysis Eilertson et al. (2012) used the generalised linear mixed model with the fixed effect part of the model as:

$$log(\mu_{ijk}) = \beta_0 + \beta^N i + \beta^D j + \beta^{ND} ij + \beta_{length} x_{ik} \quad [3]$$

where μ_{ik} is the expected number of mutations in gene k in one of the four classes indexed by i = 0,1 and j = 0,1 where i = 1 indicates nonsynonymous (N) and k = 1 divergent (D). This model estimates the intercept β_0 (the density of synonymous polymorphisms), β^N (the genome-wide difference between a mutation being nonsynonymous versus synonymous), β^D (the genome-wide difference between a mutation being a substitution versus a polymorphism), and β^{ND} (the interaction effect describing any genome-wide excess or dearth of nonsynonymous substitutions). Parameter x_{ik} is the logarithm of the number of sites in gene k where a synonymous (i = 0) or a nonsynonymous (i = 1) mutation could occur and the fixed effect β_{length} models how the number of observed mutations changes as a function of the

number of sites. Eilertson et al. (2012) also fitted a random effect structure that models between-gene mutation patterns after accounting for the fixed effects:

$$log(\mu_{ijk}) = \beta_0 + \beta^N i + \beta^D j + \beta^{ND} ij + \beta_{length} x_{ik} + \varepsilon_k + \varepsilon_k^N i + \varepsilon_k^D j + \varepsilon_k^{ND} ij \quad [4]$$

where the additional terms denoted ε are the gene-specific random deviations from each of the first four fixed effect terms described above. The four gene-specific random deviations were assumed to come from a multivariate normal distribution with estimated (co)variance matrix. Eilertson et al. (2012) define the selection effect of gene *k* as $\beta^{ND} + \varepsilon_k^{DG}$, where a positive effect is evidence for positive selection, and (in Bayesian terms) the posterior probability that the effect exceeds zero can be directly assessed.

Here we extend the *SnIPRE*-like model of Eilertson et al. (2012) to accommodate multiple species and to allow the evolutionary parameters to differ among different classes of gene. To this end we allowed the four fixed effects to vary by species and by gene class (control, piRNA, siRNA, miRNA and viRNA) to give the fixed effect model:

$$\beta_0 + \beta^N i + \beta^D j + \beta^{ND} ij + \beta_{length} x_{iklm} + \beta_{Class:l} + \beta^N_{Class:l} i + \beta^D_{Class:l} j + \beta^{ND}_{Class:l} ij + \beta_{Organism:m} + \beta^N_{Organism:m} i + \beta^D_{Organism:m} j + \beta^{ND}_{Organism:m} ij \quad [5]$$

From this we calculated the estimated selection effect for a specific pathway as $\beta^{ND} + \beta^{ND}_{Class:l}$. The random effect portion of the model included homologue-specific effects and gene-specific effects and had the form

$$u_{Hom:k} + u^N_{Hom:k} i + u^D_{Hom:k} j + u^{ND}_{Hom:k} ij + \varepsilon_{klm} + \varepsilon^N_{klm} i + \varepsilon^D_{klm} j + \varepsilon^{ND}_{klm} ij \quad [6]$$

In addition to the four gene effects, the four homologue effects were also assumed to come from a multivariate normal distribution with estimated (co)variance matrix. We used this model to calculate the selection effect for homologue *k* in gene class *l* as $\beta^{ND} + \beta^{ND}_{Class:l} + u^{ND}_{Hom:kl}$ and each gene as $\beta^{ND} + \beta^{ND}_{Class:l} + u^{ND}_{Hom:klm} + \varepsilon^{ND}_{klm}$. We estimated β_{length} rather than fixing it at one, as in Eilertson et al. (2012), although the posterior mean of β_{length} was close to one, supporting the assumption of Eilertson et al. (2012). In addition, we also fitted the SnIPRE model without assuming genes belong to known pathways, analogous to model 2. The code to fit these models is provided in the S1 text.

3.3.4 Selective sweep analysis

The recent spread of a positively selected allele leaves characteristic patterns of diversity and allele frequencies in the genomic region surrounding the selected site, and these can be used to detect recent adaptive substitutions (Smith and Haigh, 1974; Barton, 1998; Nielsen, 2005). We used SweeD (Nielsen *et al.*, 2005; Pavlidis *et al.*, 2013) to search for evidence of recent selective sweeps in the regions surrounding RNAi genes. The algorithm scans the genome and at a user-defined interval calculates the composite likelihood of the observed site frequency spectrum (SFS) under a model of a selective sweep centred on that site, versus a standard neutral model. The ratio of the two composite likelihoods (CLR) is then used as a test statistic, with significance assessed by coalescent simulation (see Figure S1). We used this method to scan 200 kb (or less if the reference genome contig was less than 200 kb) surrounding each gene of interest in each species. For each focal region, we polarised the SFS by parsimony between the outgroup reference genome and the ingroup consensus sequence, which we aligned with LastZ ungapped alignment (Harris, 2007). This simple parsimony-based inference of ancestral states risks mis-polarization of low frequency polymorphism as high-frequency derived alleles, however we assume this does not differentially affect control and RNAi genes. We did not assume an ancestral state for fixed differences that were invariant in our ingroup (i.e. these sites were folded). This will make the analysis more robust to possible errors during contig alignment, because misalignment would manifest itself as regions of increased divergence between species. We included invariant sites in the analysis, as a characteristic signature of a recent sweep is a lack of diversity, and so including invariant sites in Sweepfinder analyses can greatly improve statistical power (Nielsen *et al.*, 2005). This comes with a risk of increased false positives (Huber *et al.*, 2016), but including these sites should not differentially affect RNAi and control genes, unless there is a consistent difference in neutral mutation rates or depth of coverage between these two classes of genes. We have confirmed that there are no consistent differences in read depth between control and RNAi genes. The SweeD analysis provides CLR values for equidistant points across the genome, with CLR values forming a "peak" in areas with high support for a sweep. To assess whether RNAi genes have experienced more sweeps than control genes in 6 of our 8 species (*B. mandarina* and *P. pacificus* were not tested because the published genome assemblies are unannotated), we counted the number of RNAi and control genes that overlapped significant peaks in the CLR statistic (based on the significance threshold provided by coalescent simulation, Figure S1, S2 text). If consecutive peaks occurred within 1 kb of each

other, we classified them as a single broad peak, such that the contig was split into "sweep-positive" and "sweep-negative" areas. We then classified all genes along the contig as to whether they overlapped a "sweep-positive" area or not, and whether or not they were an RNAi gene. We used a binomial test to assess whether RNAi or control classes had more sweep-positive genes than expected given the summed gene length for each class.

To test whether sweeps were enriched in any particular subpathway, we normalised the maximum CLR statistic in a gene by the expected significance threshold from coalescent simulations and modelled these values ($\widetilde{CLR}$) using the following linear mixed model:

$$\widetilde{CLR}_{klm} = \beta_0 + \beta_{Class:l} + u_{Organism:m} + \varepsilon_{klm} \qquad [7]$$

Here, $\beta_{Class:l}$ is a fixed effect for the pathway each gene is assigned (miRNA, siRNA, piRNA or viRNA), $u_{Organism:m}$ is a random effect for species *m* and ε_{klm} is the error term.

In the four organisms for which we have haplotype information (*D. melanogaster, D. pseudoobscura, P. pacificus, C. briggsae*), we additionally tested for ongoing or soft sweeps using the haplotype-based nSL statistic (Ferrer-Admetlla *et al.*, 2014). The nSL statistic is similar to the more widely used iHS statistic (Voight *et al.*, 2006), except that distance is measured in polymorphic sites rather than the genetic map distance. This genome scan calculates the average number of consecutive polymorphisms associated with either the ancestral or derived allele at each polymorphic site along the contig across all pairwise comparisons. Areas with long range linkage disequilibrium will therefore be identified through SNPs with extreme nSL values.

3.3.5 Coalescent simulations

To assess significance in the SweeD analyses, we used ms (Hudson, 2002) to perform 1000 coalescent simulations for each gene region of interest in each species, given the observed number of segregating sites, reported recombination rate, and a previously published estimate of the demographic history of that species. When population scaled recombination rate estimates were not available, we used estimates of N_e to scale per-base rate estimates. Although the details of the demographic scenarios we modelled are unlikely to impact substantially upon our qualitative comparisons of between sweep frequency in different types of gene, we attempted to use null models consistent with the published literature. The demographic scenarios modelled for each species are illustrated in Figure S1. For *D. melanogaster*, recombination rates from the *Drosophila* recombination rate calculator were used with a

constant N_e for African populations of $1.15x10^6$ (Charlesworth, 2009). Some genes (*ael, AGO3, pasha,* and *Rm62*) are reported to lie in areas with zero recombination (Fiston-Lavier *et al.*, 2010), so we set the recombination rate in these genes at the lowest non-zero rate observed. For *D. pseudoobscura,* we simulated a population expansion (Haddrill, Loewe and Charlesworth, 2010; Larracuente and Clark, 2014), and used the population scaled rates of recombination and gene conversion from Larracuente and Clark (2014). For *Anopheles gambiae,* we used demographic history parameters from (Crawford and Lazzaro, 2010) for the Cameroon population, and the recombination rates for each individual chromosome arm (1 cM/Mb for the X, 1.3 cM/Mb for 3L and 2R, 1.6 cM/Mb for 3R, and 2 cM/Mb for 2L) from (Pombi *et al.*, 2006; Stump *et al.*, 2007). Effective population size (N_e) was set to $2.4x10^6$ estimated using the *D. melanogaster* mutation rate of (Keightley *et al.*, 2014) and the Watterson's theta (θ_W) estimate in Crawford and Lazzaro (2010). For *H. melpomene,* we simulated three Costa Rican populations corresponding to *H. melpomene, H. cydno, and H. pachinus,* using the migration rates provided in Table 2 of (Kronforst *et al.*, 2006). We used a constant recombination rate of 7.51 cM/Mb across the entire genome with an N_e of $2.1x10^6$ for *H. melpomene*, $3.3x10^6$ for cydno, and $2.7x10^6$ for *H. pachinus*. For *B. mandarina,* we modelled the "gene-flow at bottleneck" scenario (Yang *et al.*, 2014), with an N_e of 500,000 for *B. mandarina* and 73,000 for *B. mori*, and a recombination rate of 2.97 cM/Mb (Yamamoto *et al.*, 2008; Yang *et al.*, 2014). For *A. mellifera,* four subpopulations were modelled using N_e values in Table 1 of (Wallberg *et al.*, 2014), following Figure 1F in Wallberg et al (2014) when modelling past subpopulation size changes. These subpopulations share migrants, and migration rates were estimated based on F_{ST} values between subpopulations reported in (Whitfield *et al.*, 2006). A recombination rate of 19 cM/Mb is assumed to be constant across the genome (Beye *et al.*, 2006). For *C. briggsae*, coalescent simulations and SweeD analyses were carried out on the 25 "tropical" samples in order to avoid modelling complicated demographic scenarios. These are expected to have an effective population size of 60,000, and to have undergone a recent bottleneck 0.916 N_e generations in the past (Cutter *et al.*, 2006; Denver *et al.*, 2009), assuming a 60-day generation time (Barrière and Félix, 2005). We used recombination rates for *C. briggsae* from (Ross *et al.*, 2011), which are estimated to be 9.97×10^{-8} per bp per generation in autosomes and 4.6×10^{-8} per bp per generation on the X chromosome. Finally, for *P. pacificus,* four subpopulations were modelled corresponding to clade A1, A2, C, and 9 individuals whose clade was unknown (Rödelsperger *et al.*, 2014) which coalesced

0.849 N_e generations in the past (McGaughran, Morgan and Sommer, 2013). N_e was estimated by calculating θ_W for each contig and assuming a mutation rate of$2x10^{-9}$ (Weller *et al.*, 2014).To minimise differences between the real data and simulations, sites were randomly chosen to be folded, ancestrally invariant, or fixed for a derived substitution, in each case matching the numbers observed in the real data before the SweeD analysis.

3.4 Results

3.4.1 Evidence of genome-wide adaptive substitution in insects, but not nematodes

Position-matched 'control' genes (lacking RNAi-related function) allowed us to estimate the average genome-wide rate of adaptation, assuming that proximity to an RNAi gene has no effect on their rate of adaptive evolution. Our findings broadly agree with previous analyses, suggesting a substantial fraction of amino-acid substitution is adaptive across insect species (Figure 1). All insect species shared similar estimates (ω_A from 0.02 to 0.05) except for *D. pseudoobscura,* which exhibited an extremely high ω_A value of 0.16 adaptive nonsynonymous substitutions per synonymous substitution per site, 95% bootstrap interval [0.05,0.32]. Although we only sampled two nematode lineages, it is notable that both ω_A estimates were negative (*C. briggsae*: -0.20 [-0.25, -0.15]; *P. pacificus:* -0.24 [-0.27, -0.21]. This is consistent with the previously noted high ratio of nonsynonymous to synonymous polymorphism (π_A/π_S) ratio in these species, and perhaps suggests population structure and local adaptation (Rödelsperger *et al.*, 2014; Thomas *et al.*, 2015). We also calculated α, or the proportion of adaptive substitutions for each species, which reflect the same patterns observed for ω_A (Figure S2).

The cross-species SnIPRE-like model provides a formal comparison of adaptive divergence in the insect species. The structure of the model forces comparison relative to one species, for which we chose *D. melanogaster*. *Anopheles gambiae* and *Bombyx mandarina* had levels of putatively adaptive nonsynonymous divergence that were indistinguishable from those of *D. melanogaster* (MCMCp = 0.489 and MCMCp=0.616, respectively). Consistent with the DFE-alpha estimates of ω_A, *A. mellifera* and *H. melpomene* had significantly less adaptive nonsynonymous divergence than *D. melanogaster* (MCMCp = 0.04 and MCMCp < $3x10^{-4}$, respectively), whereas *D. pseudoobscura* had an increased excess of nonsynonymous divergence (MCMCp = 0.0005).

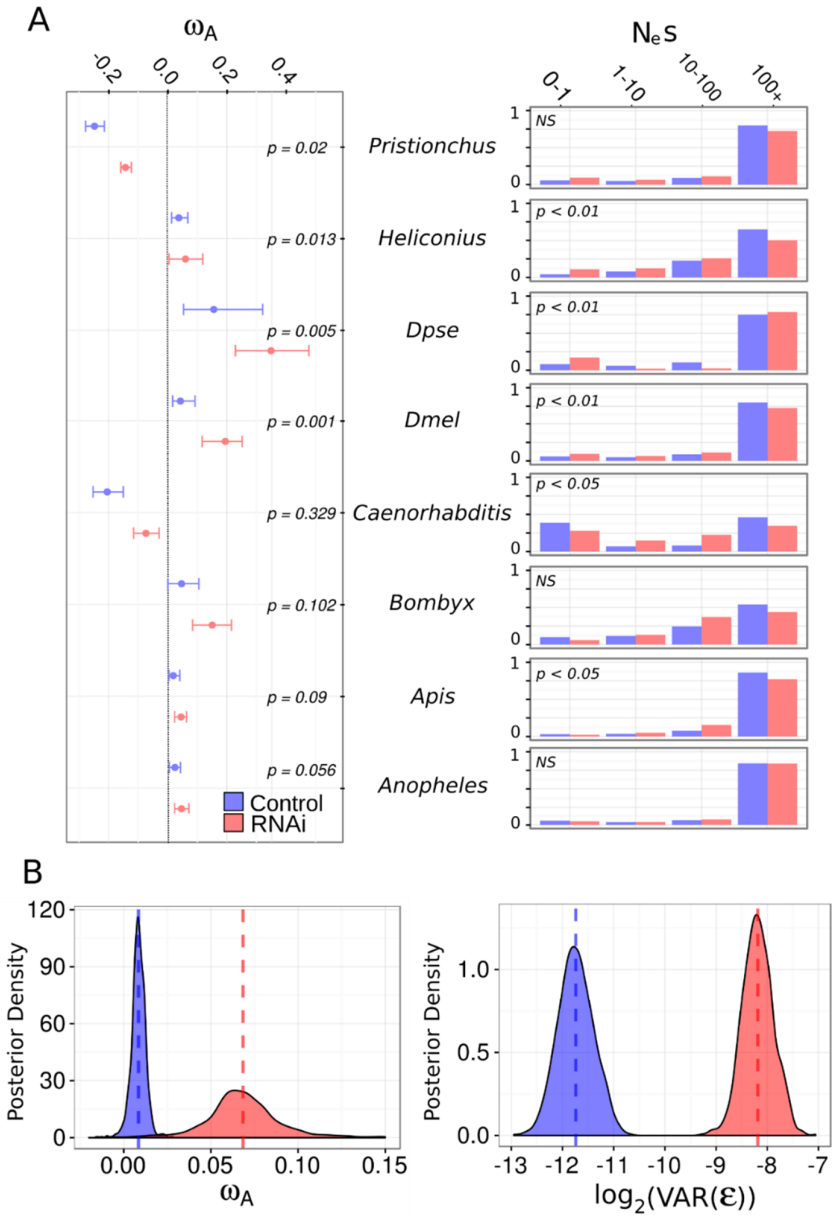

Figure 1: ω_A and the DFE differ between RNAi genes and other genes

(A) Left: For each species, ω_A estimates and 95% bootstrap confidence intervals are plotted for control (i.e. non-RNAi; blue) and RNAi (red) genes. Significance was determined by permutation. Right: The estimated discretised DFE for each species, with the proportion of mutation with deleterious $N_e s$ values in each category given for non-RNAi (blue) and RNAi (red) genes. (B) The posterior distribution of estimated ω_A for RNAi (red) versus control (blue) genes, showing that RNAi genes have much great ω_A estimates (left) and greater residual gene-level variation (right), indicating RNAi genes display higher rates adaptive amino-acid substitution but are more variable.

3.4.2 RNAi genes consistently display more adaptive protein substitution than other genes

For each focal species we estimated the distribution of fitness effects of new mutations using DFE-alpha for RNAi pathway and non-RNAi ('control') genes. We fitted two models, one in which RNAi and control genes shared a single DFE, and second in which each class of gene had a separate DFE. We then compared these models using a likelihood ratio test. In *D. melanogaster, D. pseudoobscura*, *H. melpomene, A. mellifera,* and *C. briggsae*, models in which control and RNAi genes have separate DFE parameters fitted the data significantly better than a model in which the two classes share a single DFE (Figure 1). Although there is no clear or universal trend, the DFE of control genes generally seemed slightly shifted towards more deleterious mutations than RNAi genes. For example, in most lineages (not *D. pseudoobscura* or *A. gambiae*), the estimated DFE had a higher proportion of strongly deleterious mutations in control genes than RNAi genes, which suggests less constraint in RNAi genes. However, the overall shape of the DFE is quite different between species, either indicating that in these species gene function may play a smaller role than other factors in patterns of polymorphism, such as the effective population size, or that the DFE is estimated with low precision.

We then compared rates of adaptive amino acid substitution in RNAi genes to those in the non-RNAi 'control' genes in each lineage, by pooling polymorphism and divergence data for the two classes as input to DFE-alpha (Figure 1). In every species tested, the point-estimate of class-wide ω_A was greater in RNAi genes than control genes. Although the effect was often small, the difference was individually significant in *D. melanogaster, D. pseudoobscura, H. melpomene,* and *P. pacificus.* To quantify the overall difference, we analysed individual gene estimates of ω_A in a linear mixed model framework (i.e. a meta-analysis) to estimate cross-species rates of adaptive evolution in control and RNAi genes (Figure 1). We found the cross-species ω_A was significantly greater for RNAi genes than control genes, estimated as ω_A = 0.062 [0.049, 0.078] versus ω_A = 0.01 [0.0009, 0.019] ($p < 0.001$). In addition, the residual gene-level variance was also much greater (MCMCp <0.001) for RNAi genes (0.0037, [0.0022, 0.0051]) than control genes (0.0003, [0.0001, 0.0004]), implying that ω_A is more variable in this class than among genes in general and consistent with a subset of RNAi genes or pathways undergoing extreme rates of adaptive amino acid substitution (Figure 1). However, the

coefficient of variation was not significantly different between RNAi and control genes, indicating these differences in residual variances are consistent with a mean-variance relationship in the rates of RNAi pathway genes (Figure S4).

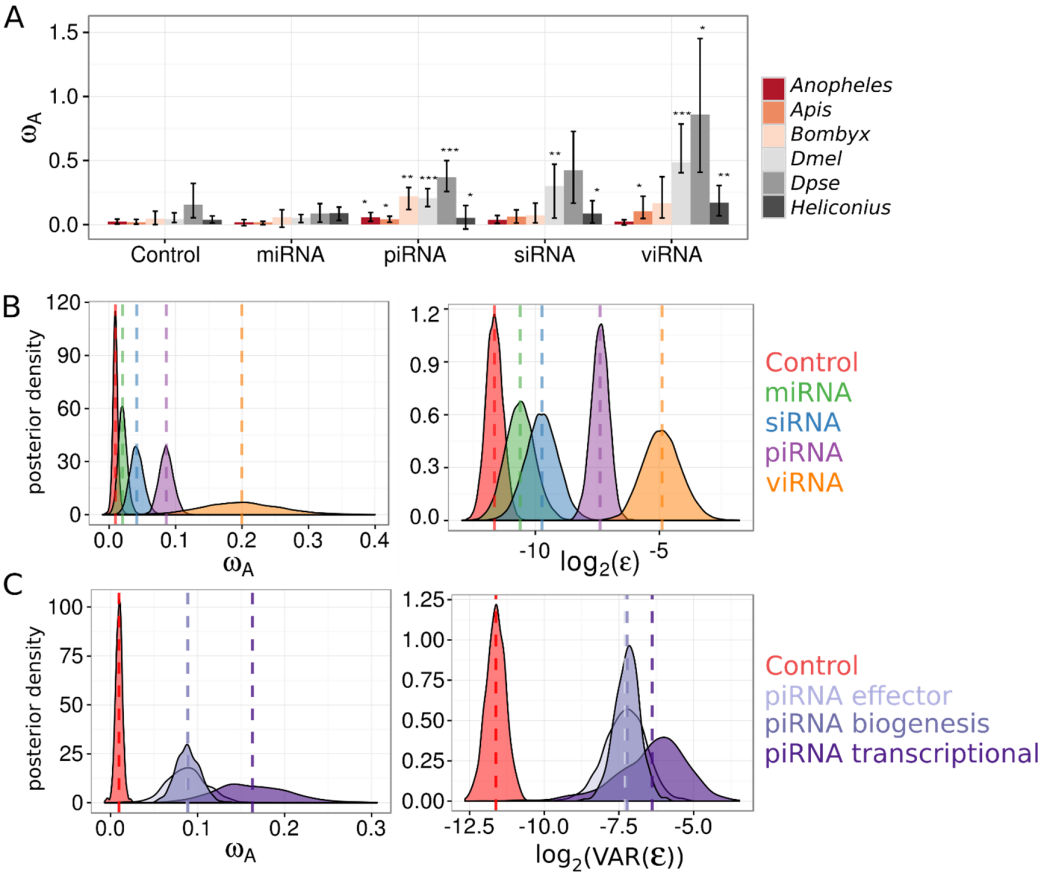

Figure 2: DFE-alpha estimates of ω_A differ among RNAi subpathways

(A) ω_A estimates from pooled polymorphism and divergence data across insect RNAi subpathways using DFE-alpha. The ω_A statistic was estimated for each subpathway in each organism and confidence intervals obtained by bootstrapping across genes. Significance was assessed by permutation tests between sub-pathway and control genes for each organism ($p < 0.05$, $p<0.01$**, $p<0.001$***). (B) Individual-gene DFE-alpha ω_A estimates were analysed using a linear mixed model in MCMCglmm (see Text S1), and show that (left) the viRNA pathway exhibits the fastest rate of adaptive protein substitution, followed by the piRNA pathway, and that among-gene variance shows the same pattern (right). (C) Individual gene DFE-alpha ω_A estimates were analysed in MCMCglmm, except that the piRNA pathway was further split into genes involved in transcriptional silencing, piRNA biogenesis, or piRNA-mediated effectors of silencing. The posterior distributions of these three effect sizes versus control genes are plotted. All three piRNA functions are targets of elevated positive selection and have large residual variances, although genes mediating transcriptional silencing have greater point estimates for both.*

3.4.3 Adaptive rates are high in piRNA and viRNA pathways

The higher rate of adaptive substitution seen in RNAi genes as a whole could result from slightly elevated positive selection across all components, or to a subset of the genes or pathway being substantially elevated. The higher gene-level variance seen in RNAi genes (above) suggests the latter, and to test this we pooled polymorphism and divergence data by sub-pathway for each insect species to calculate rates of adaptation in miRNA, siRNA, viRNA (i.e. confirmed antiviral siRNA in *D. melanogaster*), and piRNA pathways (Figure 2). In each species, the piRNA pathway exhibited a significantly greater rate of adaptive amino acid substitution than control genes, and miRNA pathway genes showed similar rates to control genes. Rates of adaptation for the siRNA (both endo-siRNA and viRNA) pathway were greater in only a subset of lineages. The magnitude of rates and proportion of lineages nominally significant in the test increased upon removing endo-siRNA genes and restricting the analysis to viRNA genes only. For all subsequent analyses, we analysed these pathways separately to test the hypothesis that the core antiviral RNAi genes have elevated rates of adaptive evolution.

To formalise the effect of pathway (miRNA, piRNA, non-antiviral endo-siRNA, viRNA) while accounting for variability in adaptation across species (Figure 2), we performed a meta-analysis of ω_A estimates in individual genes from DFE-alpha, fitting pathway as a fixed effect. The piRNA, viRNA, and endo-siRNA pathways were each significantly different from control genes (control ω_A =0.01 [0.002,0.018]; piRNA MCMCp < 0.001; viRNA MCMCp = 0.002; siRNA MCMCp = 0.004; for MCMCp value calculation, see the S1 text), with cross-species estimates of ω_A of 0.08 [0.06,0.10], 0.18 [0.06, 0.30] and 0.03 [0.01,0.05], respectively. The viRNA pathway ω_A estimate was not significantly greater than the piRNA pathway (MCMCp = 0.07), but was greater than the endo-siRNA pathway (MCMCp = 0.01), and the miRNA pathway (MCMCp < 0.001). The ω_A estimate for the piRNA pathway was significantly greater than the endo-siRNA (MCMCp = 0.002) and the miRNA pathways (MCMCp < 0.001). Consistent with our analysis of pooled polymorphism and divergence data, the rate of adaptive evolution in the miRNA pathway (ω_A = 0.01 [-0.001, 0.02]; MCMCp=0.09) was not significantly different from control genes. Our linear models included pathway-specific error variances, which were lower for control genes (3 [2,4] $x10^{-4}$) and miRNA pathway genes (7 [2,12] $x10^{-4}$) than for endo-siRNA (13 [4,22] $x10^{-4}$), piRNA (66 [37,97] $x10^{-4}$), and viRNA pathway genes (0.04 [0.007, 0.86]), consistent with a great variation in adaptive rates in these pathways. As in the comparison between RNAi and control genes, these elevated variances in piRNA, siRNA, and

viRNA pathways could be explained by the elevated mean rates in these pathways (Figure S4).

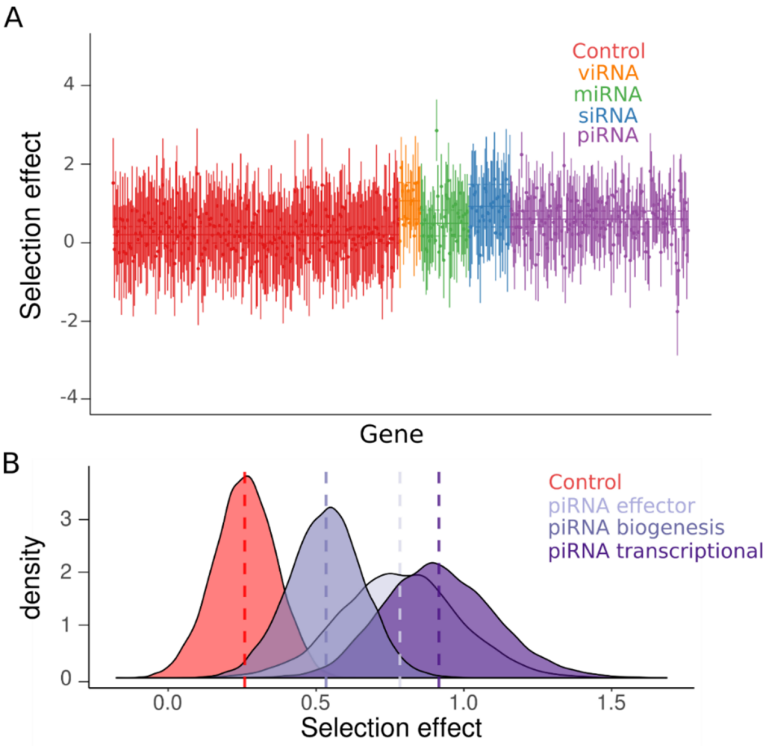

Figure 3 SnIPRE-like selection effects

(A) SnIPRE 'selection effect' with 95% confidence intervals (species-level effects removed) are plotted for each gene in each species, coloured according to the gene's role in the RNAi pathway. Solid horizontal lines signify the mean selection effect for each RNAi subpathway (or control genes) with dotted lines signifying the 95% confidence intervals for the subpathway mean. SnIPRE and DFE-alpha analyses are consistent in suggesting that the viRNA, endo-siRNA, and piRNA pathway have more adaptive amino-acid substitutions than control genes. The largest selection effect was seen in the Bombyx mandarina Dcr-1 locus, with a selection effect of 2.95 (Figure S6). (B) We also performed a SnIPRE analysis after dividing the piRNA pathway into three functional classes, as in Figure 2. The posterior distribution of selection effects associated with each piRNA function are plotted. Similar to DFE-alpha, SnIPRE identifies all three pathways as significantly elevated relative to control genes, however in the SnIPRE analysis transcriptional silencing genes have a significantly greater adaptive rate than biogenesis factors.

We repeated the subpathway-level analysis using a SnIPRE-like model (Eilertson, Booth and Bustamante, 2012) to estimate the average selection effect within subpathways across organisms without making any explicit assumptions about the DFE. Although SnIPRE can be used to provide estimates of population genetic parameters, we limit our discussion to the "selection effect" statistic, where negative values are consistent with constraint and positive values with adaptive protein evolution, and magnitude reflects the strength of positive or negative selection. Consistent with our analysis of DFE-alpha estimates, the SnIPRE model

identified a mean positive selective effect estimated across species (selective effect=0.25 [0.02, 0.46], MCMCp = 0.03), with large variance among genes (Figure 3). Again, viRNA, endo-siRNA, and piRNA pathway-level selection effects were significantly elevated compared to control genes (viRNA: 1.10 [0.63, 1.57] MCMCp < $5x10^{-4}$, non-antiviral siRNA: 0.96 [0.44, 1.52] MCMCp = 0.02, piRNA: 0.63 [0.44, 0.84] MCMCp < $3x10^{-4}$), with the viRNA pathway exhibiting a significantly larger effect than the piRNA (MCMCp = 0.006), but not the endo-siRNA (MCMCp = 0.66). In agreement with the DFE alpha analysis, the miRNA pathway was not significantly different from control genes (MCMCp = 0.07), and had a selection effect of 0.53 [0.20, 0.86].

3.4.4 Adaptation is elevated in all major piRNA pathway functions, but is most enriched in transcriptional silencing

Rapid adaptation in *Drosophila* piRNA pathway genes has been hypothesized to be the result of fluctuating selection for increased TE defence and decreased off-target genic silencing (Blumenstiel, Erwin and Hemmer, 2016). A prediction of this hypothesis is that genes involved in transcriptional silencing would be under increased positive selection. We tested this prediction by further dividing the piRNA pathway into effectors (e.g. PIWIs), biogenesis factors (e.g. adapter proteins), and transcriptional silencing factors, and using single-gene polymorphism and divergence data to estimate ω_A and the selection effect for each piRNA functional category. We found all piRNA functional groups are significantly greater than control genes (MCMCp < 0.001) (Figure 2C), and that transcriptional silencing genes (ω_A = 0.16 [0.08-0.25]) have greater adaptive rates than effectors (MCMCp = 0.04, ω_A = 0.08 [0.04-0.13]) and biogenesis factors (MCMCp = 0.03, ω_A = 0.08 [0.05-0.11]). This result holds when excluding *Drosophila* transcriptional silencing factors *rhino*, *deadlock*, and *cutoff*, which are products of recent gene duplication or *de novo* formation (Figure S3), and may not have evolutionary rates that are directly comparable to other genes.

We also estimated the average selection effect for each functional process of the piRNA pathway using the SnIPRE approach. Similar to the DFE-alpha meta-analysis, we find that all piRNA functional categories have elevated positive selection relative to control genes (biogenesis: MCMCp=0.018, effector: MCMCp=0.012, transcriptional silencing: MCMCp=0.0004), that transcriptional silencing factors had the largest average selection effect of 0.92 [0.58, 1.31], and that genes involved in transcriptional silencing were significantly greater than biogenesis factors (selection effect: 0.53, [0.29, 0.78], MCMCp = 0.027) (Figure 3B). In contrast to the

DFE-alpha meta-analysis, however, genes involved in transcriptional silencing were not significantly greater than effector genes (0.78 [0.40, 1.19], MCMCp = 0.68), and pathway-level point estimates of these selection effects were much closer (Figure 2C, Figure 3B).

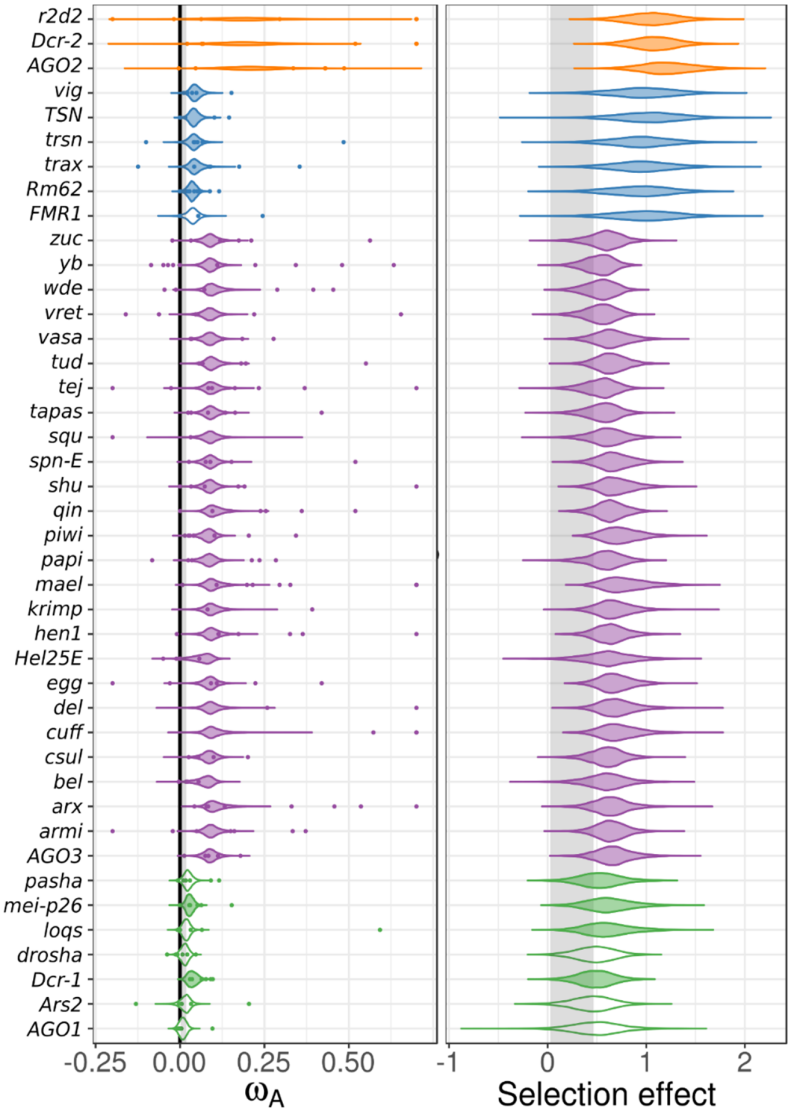

Figure 4 Cross-species homologue-level estimates of ω_A and selection effects

(Left) Individual homologue ω_A estimates (coloured points) were calculated using DFE-alpha and analysed using a linear mixed model with subpathway as fixed effect and species and homologue as a random effect (estimate uncertainty was included by incorporating bootstrap intervals as measurement error variance). The posterior distributions of the cross-species estimate for ω_A for each homologue are plotted, and shaded if significantly different from the control gene distribution (region shaded grey). Single-gene estimates of $\omega_A > 0.75$ are plotted at 0.75 for clarity. (Right) The analogous analysis performed using SnIPRE, with the posterior distribution of homologue-level selection effects plotted. Both analyses find little variation among homologues after accounting for subpathway, and homologue-level analyses generally mirror pathway-specific analyses. See Figure S4 for the equivalent models that exclude the fixed effect of pathway.

3.4.5 Individual genes in the piRNA and viRNA pathway show elevated adaptation

The higher overall rates of adaptive protein substitution seen in RNAi genes may result from the engagement of some genes in an evolutionary arms race (e.g. with viral suppressors of RNAi), a response to the selection imposed by the invasion of novel parasites (e.g. transposable elements), or a trade-off between the specificity and sensitivity of genome defense (Obbard *et al.*, 2006; Aravin, Hannon and Brennecke, 2007; Blumenstiel, Erwin and Hemmer, 2016). We used a linear mixed model to combine single-gene estimates of ω_A from DFE-alpha across multiple species to identify candidate arms race genes in the RNAi pathways, fitting subpathway as a fixed effect, with homologue and organism as random effects, and subpathway-specific error variances. We found little variation among genes in a subpathway after accounting for subpathway, and in most cases there was not enough information to differentiate individual genes from the subpathway mean (Figure 4, left). Consequently, all genes in the rapidly evolving viRNA and piRNA subpathways were identified as having significantly greater adaptive rates than control genes. Most (5 of 6) siRNA pathway genes, and 2 of 7 miRNA pathway genes were also identified as having significantly elevated adaptive rates. Although a model that accounts for pathway is statistically preferable if pathways are meaningful, any errors in assigning 'pathway' membership would introduce bias to the estimates for misclassified genes. We therefore also estimated homologue-specific effects in a model that excludes the subpathway effect. This model finds significant evidence for positive selection in fewer genes (Figure S5, left) including 13 of 22 piRNA genes, 2 of 3 viRNA genes, and no genes in the siRNA or miRNA pathway.

We also performed this homologue-level analyses using the SnIPRE approach. Similar to the DFE-alpha meta-analysis, we found very little information after accounting for subpathway (Figure 4, right), resulting in low among-gene variation within RNAi subpathways. When we excluded subpathway effects, we found a similar result to the homologue-level DFE-alpha meta-analysis without subpathway, except fewer piRNA pathway genes are nominally significant (6 of 22 genes) (Figure S5, right). Notably, *maelstrom, eggless, piwi* (incorporating the dipteran duplicate *aub*)*, AGO2,* and *Dcr-2* were found to have significantly elevated positive selection across all four homologue-level analyses (i.e. with or without imposing a subpathway classification).

MK tests are commonly used to test for positive selection in individual genes. SnIPRE selection effects can be used to perform an analogous test for selection, except the approach can gain power by taking in the genome-wide distribution of polymorphism and divergence patterns by fitting gene as a random effect (Eilertson, Booth and Bustamante, 2012). We found that 36% of RNAi genes show nominally 'significant' evidence for adaptive protein evolution across species analysed. In contrast, only 5% of selection effects in control genes were significantly positive (Figure S6). At the pathway level, 40% of piRNA genes, 44% of viRNA genes, 26% of non-antiviral siRNA pathway genes, and 25% of miRNA pathway genes had significantly positive selection effects (Figure S6). No gene had positive selection effects in every lineage, although *armitage, capsuleen, cutoff, tudor, vasa, vretano,* and *Yb* homologs were identified in over half the lineages.

3.4.6 Selective sweeps are detectable across functional classes of RNAi genes

Recent positive selection is expected to leave a characteristic mark in the genome, including a SFS skewed towards low and high frequency alleles and a local reduction in polymorphism (Smith and Haigh, 1974; Barton, 1998; Nielsen, 2005). As RNAi genes show elevated rates of adaptive evolution, we speculated that they may also exhibit more evidence of recent selective sweeps. Using SweeD, we found that many of the insect lineages do show evidence for sweeps in a subset of RNAi genes (Figure 5, Figures S7-S14). We tested whether RNAi genes have undergone more recent sweeps than surrounding genes by classifying nominally significant peaks as either occurring near (within 1 KB) an RNAi gene or not, and using a binomial test to determine whether more sweeps than expected occur in RNAi genes (given their length). In four of the six species tested (*D. melanogaster, D. pseudoobscura, A. mellifera,* and *A. gambiae*) there were significantly more detectable sweep signals in RNAi genes than in surrounding non-RNAi genes (*D. melanogaster p* = 0.0006; *A. mellifera p* = 0.015*; A. gambiae p* = 0.0001*; D. pseudoobscura p* = 7×10^{-5}). However, we find no difference among subpathways in the frequency with which we detected recent sweeps. General differences in constraint between RNAi and control genes could bias these results, perhaps through misidentification of reduced diversity caused by elevated negative or background selection as a sweep. However, DFE-alpha and SnIPRE analyses suggest RNAi genes are less constrained (Figure 1; Text S1, model 3A, nonsynonymous:piRNA effect), making our analysis conservative. None of the genes exhibited a significant CLR peak across all organisms tested, although *spn-E* and *vig* display significant evidence of recent sweeps in five of the six insect lineages.

It was notable that 34% of the variation in the per-gene maximum CLR test statistic was attributable to species, consistent with either sample size or demographic history playing a substantial role in our power to detect sweeps.

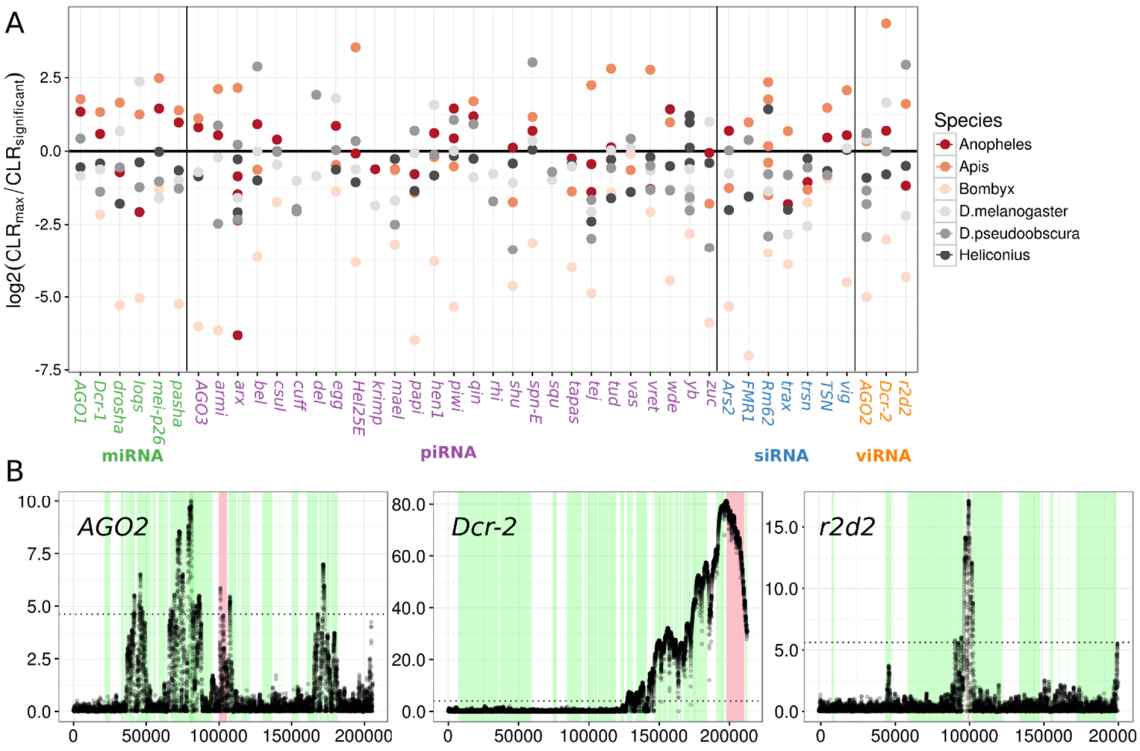

Figure 5 Selective sweeps in RNAi genes and example SweeD plots

(A) Points indicate the $\log_2$ ratio of the maximum observed CLR value (from SweeD) in the named gene to the CLR 95% significance threshold inferred from simulation. Values above 0 indicate there was a 'significant' CLR peak in a genic region and colours indicate species. (B) The viRNA pathway in Apis mellifera shows strong evidence for recent sweeps. For each of the three viRNA pathway genes the CLR statistic is plotted across a 200 kb region. The dotted line is the significance threshold estimated through neutral simulations under a published demographic history. Red regions denote the focal gene and green regions highlight surrounding genes. In Apis, both Dcr2 and R2D2 show strong evidence for sweeps with the surrounding region of Dcr2 being devoid of polymorphism, indicating this sweep was recent and rapid. AGO2 also shows a significant peak, but this is narrow and only marginally significant.

Sweep signatures were the most pronounced in *A. mellifera*, in both the CLR magnitude and breadth of the genomic region affected (Figure 5, Figure S12). These were associated with large regions devoid of any polymorphism, despite the high rate of recombination seen in honeybees (Beye *et al.*, 2006), which is expected to narrow the region affected by a nearby

sweep. We also searched for evidence of haplotype structure, as would be expected during an ongoing or soft selective sweeps using the nSL statistic (data not shown). However, there were no strong signals in any of the RNAi genes for which we had haplotype information.

3.5 Discussion

Using both DFE-alpha and SnIPRE-like McDonald-Kreitman analyses we identify elevated rates of adaptive evolution in RNAi-pathway genes across six insects and two nematodes. In most species, the RNAi-pathway genes were also more likely to display evidence of a recent selective sweep. As in *Drosophila*, genes involved in the suppression of viruses and transposable elements show the highest rates of adaptive evolution, consistent with these genes being engaged in an arms race in multiple invertebrate lineages. We were able to extend past *Drosophila* analyses by combining genic rates of adaptive evolution across species to infer positive selection associated with particular RNAi pathway functions and homologues. We found accelerated adaptation across piRNA pathway functions, including piRNA biogenesis machinery, effector proteins, and especially transcriptional silencing machinery. Although there was substantial variation in rates among RNAi genes, the antiviral genes *AGO2* and *Dcr-2* and the piRNA pathway genes *maelstrom, eggless, piwi*, *aub, armitage, capsuleen, cutoff, tudor, vasa, vretano, spn-E, vig* and *Yb* show consistently strong signatures of positive selection.

3.5.1 Identification of rapidly evolving pathways by DFE-alpha and SnIPRE

Estimated rates of adaptive protein evolution in an MK-framework (McDonald and Kreitman, 1991) can be biased by past population size changes and slightly deleterious mutations that segregate at low frequencies (McDonald and Kreitman, 1991; Eyre-Walker, 2002; Fay, Wyckoff and Wu, 2002; Smith and Eyre-Walker, 2002; Sawyer *et al.*, 2003; Bierne and Eyre-Walker, 2004; Welch, 2006; Charlesworth and Eyre-Walker, 2008; Eyre-Walker and Keightley, 2009; Gossmann, Keightley and Eyre-Walker, 2012). Here we attempted to account for these biases by explicitly modelling the DFE and demographic history using DFE-alpha (Eyre-Walker and Keightley, 2009), or by modelling the genome-wide patterns of polymorphism and divergence with SnIPRE (Eilertson, Booth and Bustamante, 2012). We expect these methods to complement one another. Both methods assume that demographic history affects all loci in a similar manner, but SnIPRE is better able to capture variability across loci in these effects, while DFE-alpha conditions on a point estimate. In addition, both methods can be biased by

variation in the DFE, but under different circumstances. The DFE-alpha meta-analysis assumes a similar DFE for all genes, which may bias single-gene ω_A estimates either direction depending on the true DFE of a gene. SnIPRE, like traditional MK-style analyses, does not take into account the site frequency spectrum, and so can be biased by slightly deleterious mutations which will downwardly biased selection effects. Therefore, for single-gene analyses SnIPRE is more conservative, while DFE-alpha can be more powerful if the SFS and DFE parameters are accurately estimated.

Most of the qualitative results of each of these analyses agree that genes in the piRNA and viRNA pathways are evolving adaptively. However, SnIPRE and DFE-alpha analyses disagree on the relative differences in the rate of adaptive evolution among subpathways. For example, the DFE-alpha meta-analysis provides low point estimates for the endo-siRNA and miRNA pathways relative to the piRNA and viRNA, but SnIPRE identifies the endo-siRNA selection effect as higher than the piRNA, and piRNA genes closer to the miRNA. As noted above, this incongruence could reflect differences in the DFE between subpathways; genes in the miRNA and endo-siRNA pathways are highly conserved and have low rates of protein evolution, while mechanisms of piRNA pathway function are surprisingly diverse across animals (Morazzani *et al.*, 2012; Sarkies *et al.*, 2015). These differences in constraint could lead to an underestimation of miRNA and endo-siRNA pathway adaptation and overestimation of piRNA adaptation in the DFE-alpha analyses, and indicate that estimating the DFE separately for each subpathway may improve estimates.

3.5.2 Adaptive protein evolution across species is enriched in specific functional pathways

We found large differences in rates of adaptive protein substitution between insects and nematodes, but less variation among insect species. In an analysis of variance, we find that species explained only 11% of the variation in gene-level estimates of ω_A but, gene and pathway explained 42% of the variation in gene-level ω_A estimates, suggesting that gene function is a greater determinant of the rate of adaptive evolution than species. The elevated rate seen in piRNA and viRNA pathway genes across species could be caused by rapid adaptation in the same subset of genes in a pathway, or in a random selection of genes in a pathway. Homologue-level analysis of ω_A and selection effects (Figure 4, Figure S5) indicates it is probably both, as subsets of homologues within pathways show consistent evidence for elevated

adaptive protein evolution, but homologous genes also exhibit high variances across species (but see Figure S4).

Much of the variation in adaptive rate is not attributable to species or conserved gene function, and it is necessarily difficult to ascribe this remaining variance to a source. It is likely that the great majority is derived from the sampling error associated with measuring polymorphism and divergence in a single gene, however biological processes may also contribute. Functional divergence of a gene from its role in *Drosophila* could affect the adaptive rate in that species. For example, the repurposing of the piRNA pathway to target viruses in mosquitoes might be expected to increase adaptive rates of any factors shared in both anti-TE and antiviral roles (Morazzani *et al.*, 2012). Additionally, if conflict is driving the observed adaptation, then differences in the magnitude or frequency of conflict could change the adaptive potential of a gene. In nature, this could be driven by differences in the diversity, frequency, or virulence of viral pathogens across species.

3.5.3 Potential Drivers of Adaptation in the viRNA pathway

It seems likely that the elevated rates of adaptive protein evolution we detect in the viRNA and piRNA pathways are a result of recurrent selection mediated by viruses and/or TEs. First, it is well established that defensive pathways show high rates of adaptive evolution, presumably as a consequence of antagonistic coevolution with parasites (Stenseth and Smith, 1984; Buckling and Rainey, 2002; Paterson *et al.*, 2010; Brockhurst *et al.*, 2014). For example, a recent analysis of virus-interacting proteins estimated that 30% of adaptive protein changes in mammals are driven by viruses (Enard *et al.*, 2016). Second, for the viRNA pathway genes at least, viral suppressors of RNAi are strong candidates to be the driving agent. Many RNA and DNA viruses of invertebrates are known to have proteins or structural RNAs which actively block RNAi function (Li, Li and Ding, 2002; Van Rij *et al.*, 2006; Nayak *et al.*, 2010; Joël T. van Mierlo *et al.*, 2012; Alfred W Bronkhorst *et al.*, 2014), and these can evolve rapidly and can be highly host-specific, consistent with an arms-race scenario (van Mierlo *et al.*, 2014). We find that *AGO2* and *Dcr-2* display consistently elevated rates of adaptive protein substitution across insect species, with additional limited evidence of elevated adaptation in *hen1*, all of which have previously been identified as targets of active suppression by viral proteins (viral suppressors of RNAi; VSRs)(Van Rij *et al.*, 2006; Vogler *et al.*, 2007; Nayak *et al.*, 2010; Joël T. van Mierlo *et al.*, 2012; Van Cleef *et al.*, 2014), lending credibility to the hypothesis that viruses may play a major role in driving the observed rapid evolution in RNAi genes.

3.5.4 Potential Drivers of Adaptation in the piRNA pathway

Whereas an arms-race between antiviral RNAi genes and viral suppressors of RNAi is intuitive, the observed rapid adaptive evolution of piRNA pathway genes is currently harder to explain. Similar to viruses, TEs are costly for their hosts and could in principle select for increased suppression (Charlesworth, Sniegowski and Stephan, 1994). However, piRNA-generating clusters ostensibly provide an adaptive defence that can arise on much shorter time scales than fixation of advantageous mutations, reminiscent of acquired immunity (Brennecke *et al.*, 2007; Khurana *et al.*, 2011; Han *et al.*, 2015; Mohn, Handler and Brennecke, 2015).

The adaptive response in piRNA genes could be mediated by at least three non-exclusive mechanisms: (i) direct piRNA pathway suppression by TEs or by off-target VSRs, (ii) recurrent "retuning" of piRNA machinery after a novel TE invasion (Lee and Langley, 2012; Yi *et al.*, 2014), or (iii) fluctuating selection on the sensitivity to detect transposon sequences and specificity to exclude off-target genic silencing (i.e. the "genomic auto-immune hypothesis") (Blumenstiel, Erwin and Hemmer, 2016). Besides the global de-repression of transposons upon invasion of the Penelope retroelement in *D. virilis* (Petrov *et al.*, 1995; Evgen'ev *et al.*, 1997; Rozhkov *et al.*, 2010; Blumenstiel, Erwin and Hemmer, 2016), there is limited evidence for (i), and the mechanism underlying this phenomenon still awaits elucidation. The latter two hypotheses are not mutually exclusive, and both posit that piRNA adaptation occurs in response to recurrent horizontal transfer of new TEs into the genome, a common occurrence in insects (Peccoud *et al.*, 2017). In (ii), the piRNA pathway evolves to optimise defence against the current suite of transposons, becoming "less adapted" for dealing with historic, obsolete ones. This would result in a Red Queen-like scenario, but instead of antagonistic coevolution with one parasite, the piRNA pathway must defend against a constant recycling of TE lineages. As the germline cells face a higher TE diversity than somatic tissues, this is broadly supported by our observation that piRNA pathway genes with primarily germline function (Czech *et al.*, 2013; Handler *et al.*, 2013; Muerdter *et al.*, 2013) have higher rates of adaptive protein evolution than those functioning in the somatic layer of cells surrounding the *Drosophila* ovary (Figure S15), although this is difficult to disentangle from previous observations that germline-specific genes have generally high adaptive rates (Choi and Aquadro, 2015; Flores *et al.*, 2015). The genomic autoimmunity hypothesis (iii) goes further, and proposes piRNA pathway adaptation to TE invasions results in increased piRNA function

and associated off-target genic effects, which are then selected against after the TE is supressed (Blumenstiel, Erwin and Hemmer, 2016). It could be argued that our analysis of adaptive rates in piRNA functions lends broad support for this, in that genes mediating transcriptional silencing show the greatest adaptive rates across species in the piRNA pathway, with additional evidence for rapid adaptation in biogenesis factors, whose rates are expected to be correlated with the transcriptional machinery (Blumenstiel, Erwin and Hemmer, 2016). However, our pathway-level and homologue-level analyses also find signals of elevated adaptation in effector genes, which have rates that covary to a lesser degree with other piRNA factors (Blumenstiel, Erwin and Hemmer, 2016). This does not refute the genomic autoimmunity hypothesis, but may suggest additional selective forces acting on the piRNA pathway independent of genes underlying a trade-off between sensitivity and specificity. Nevertheless, our results would also fit within the context of (ii), in a scenario where the transcriptional machinery has a greater evolutionary potential than the rest of the piRNA pathway.

3.5.5 Concluding remarks

Accelerated adaptive evolution in RNAi genes has been described in multiple *Drosophila* species, where a subset of genes evolve adaptively in siRNA and piRNA pathways, but not the miRNA pathway. Our analyses extend the observation of rapid RNAi gene evolution, and generalise elevated positive selection in piRNA and viRNA pathways across 6 insect and two nematode species.

Chapter 4: Isolation of a *Drosophila* DNA virus and characterization of resistance and immune responses

The text of this chapter is published:

Palmer, W. H., Medd, N. C., Beard, P. M., & Obbard, D. J. Isolation of a natural DNA virus of Drosophila melanogaster, and characterisation of host resistance and immune responses. PLoS pathogens, 14(6), (2018) e1007050.

I wrote this chapter with comments and text edits from Darren Obbard and Pip Beard. Nathan Medd made the RNA-sequencing libraries used during this study.

In addition to the supplemental figures provided in the appendix of this chapter, supplemental tables and code are available via Figshare (DOI: 10.6084/m9.figshare.c.3936037.v1).

4.1 Abstract

Drosophila melanogaster has played a key role in our understanding of invertebrate immunity. However, both functional and evolutionary studies of host-virus interaction in *Drosophila* have been limited by a dearth of native virus isolates. In particular, despite a long history of virus research, DNA viruses of *D. melanogaster* have only recently been described, and none have been available for experimental study. Here we report the isolation and comprehensive characterisation of Kallithea virus, a large double-stranded DNA virus, and the first DNA virus to have been reported from wild populations of *D. melanogaster*. We find that Kallithea virus infection is costly for adult flies, reaching high titres in both sexes and disproportionately reducing survival in males, and movement and late fecundity in females. Using the *Drosophila* Genetic Reference Panel, we quantify host genetic variance for virus-induced mortality and viral titre and identify candidate host genes that may underlie this variation, including *Cdc42-interacting protein 4*. Using full transcriptome sequencing of infected males and females, we examine the transcriptional response of flies to Kallithea virus infection and describe differential regulation of virus-responsive genes. This work establishes Kallithea virus as a new tractable model to study the natural interaction between *D. melanogaster* and DNA viruses, and we hope it will serve as a basis for future studies of immune responses to DNA viruses in insects.

4.2 Introduction

Studies of *Drosophila melanogaster* are central to our understanding of infection and immunity in insects. Moreover, many components of the *Drosophila* immune response, including parts of the JAK-STAT, IMD, and Toll (and perhaps RNA interference; RNAi) pathways are conserved from flies to mammals (Dupuis *et al.*, 2003; Sharma *et al.*, 2003; Dostert *et al.*, 2005; Zambon *et al.*, 2005; Wang *et al.*, 2006; Avadhanula *et al.*, 2009; Li *et al.*, 2013; Maillard *et al.*, 2013), making *Drosophila* a valuable model beyond the insects. The experimental dissection of antiviral immune pathways in *Drosophila* has benefited from both natural infectious agents of *Drosophila*, such as Drosophila C Virus (DCV) and Sigma virus (DmelSV), and from artificial infections, such as Cricket paralysis virus (isolated from a cricket), Flock House Virus (from a beetle), Sindbis virus (from a mosquito) and Invertebrate Iridescent Virus 6 (from a moth). However, while the availability of experimentally tractable, but non-natural, model viruses has been a boon to studies of infection, it also has two potential disadvantages. First, the coevolutionary process means that pairs of hosts and pathogens that share a history may interact very differently to naive pairs (Ferguson and Read, 2002; Compton, Hirsch and Emerman, 2012). For example, the Nora virus of *D. immigrans* expresses a viral suppressor of RNAi that is functional in the natural host, but not in *D. melanogaster* (van Mierlo *et al.*, 2014). Second, if our aim is to understand the coevolutionary process itself, then the standing diversity in both host and virus populations may be fundamentally altered in coevolving as opposed to naïve pairs. For example, heritable variation for host resistance was detectable for two natural viruses of *D. melanogaster*, but not for two non-natural viruses (Magwire *et al.*, 2012). This difference was in part due to large-effect segregating polymorphisms for resistance to the natural viruses, which are predicted to result from active coevolutionary dynamics (Contamine, Petitjean and Ashburner, 1989; Magwire *et al.*, 2011, 2012; Cogni *et al.*, 2016).

Experimental studies of host-virus interaction using *Drosophila* have consequently been limited by a lack of diverse natural virus isolates. In particular, no natural DNA viral pathogens of *D. melanogaster* have previously been isolated (Brun and Plus, 1980; Huszar and Imler, 2008; Unckless, 2011), and all natural (and most artificial) studies of viral infection in *D. melanogaster* have therefore focussed on the biology of RNA viruses and resistance to them (Xu and Cherry, 2014). For DNA viruses, our molecular understanding of insect-virus interaction has instead been largely shaped by the response of lepidopterans to their natural

baculoviruses. These are often of agronomic and/or ecological importance (Herniou *et al.*, 2004), but lack the genetic toolkit of *D. melanogaster*. Nevertheless, Lepidopteran studies of the expression response to baculovirus infection have implicated host genes with a diverse array of functions, including cuticle proteins, reverse transcriptases, and apoptotic factors, suggesting previously uncharacterised and/or host-specific antiviral immune mechanisms (Breitenbach, Shelby and Popham, 2011; Nguyen, Nielsen and Reid, 2013; Noland *et al.*, 2013; McTaggart *et al.*, 2015).

To date, the only DNA virus studies in *D. melanogaster* have used Insect Iridescent Virus 6 (IIV6), an enveloped dsDNA moth iridovirus with a broad host range (Williams, 2008). This work has shown that *Drosophila* RNAi mutants are hyper-susceptible to IIV6 infection, and that IIV6 encodes a viral suppressor of RNAi, indicating that at least some immune responses to DNA viruses overlap with those to RNA viruses (Bronkhorst *et al.*, 2012; Kemp *et al.*, 2013; Alfred W Bronkhorst *et al.*, 2014). However, while IIV6 injections are lethal in *D. melanogaster*, and IIV6 has provided useful information about the *Drosophila* response to DNA viruses, for the reasons described above it is hard to interpret the implications of this for our understanding of natural host-virus interaction.

Metagenomic sequencing has recently identified several natural dsDNA nudivirus infections in wild-caught *Drosophila*, including in *D. innubila* (*D. innubila* Nudivirus, DiNV) and in *D. melanogaster* and *D. simulans* ('Kallithea virus', KV; 'Esparto virus' (KY608910.1), and 'Tomelloso virus' (KY457233.1)), and also ssDNA densovirus infections in *D. melanogaster* and *D. simulans* ('Vesanto virus' (KX648534.1), 'Linvill Road virus' (KX648536.1), and 'Viltain virus' (KX648535.1)) (Unckless, 2011; Webster *et al.*, 2015; Hill and Unckless, 2017b; Kapun *et al.*, 2018). Like other members of the *Nudiviridae*, DiNV and KV are enveloped dsDNA viruses of around 120-230Kbp with 100-150 genes. This recently-recognised family forms a clade that is either sister to, or paraphyletic with, the Bracoviruses (Thézé *et al.*, 2011) that have been 'domesticated' by Braconid parasitoid wasps following genomic integration, and now provide essential components of the wasp venom (Herniou *et al.*, 2013; Gauthier, Drezen and Herniou, 2017). Together, the nudiviruses and bracoviruses are sister to the baculoviruses, which are arguably the best-studied dsDNA viruses of insects. They share many of their core genes with baculoviruses, but canonically lack occlusion bodies (Wang and Jehle, 2009). PCR surveys of wild flies suggest that DiNV is common in several species in the subgenus *Drosophila*, and that KV is widespread and common in *D. melanogaster* and *D. simulans*,

being detectable in 10 of 17 tested populations, with an estimated global prevalence of 2-7% (Webster *et al.*, 2015). However, we currently know little about the interaction between these viruses and their hosts. Indeed, although studies of wild-caught *D. innubila* individuals infected by DiNV suggest that infection is costly (Unckless, 2011), in the absence of an experimental *D. melanogaster* nudivirus isolate, it has not been possible to capitalise the power of *D. melanogaster* genetics to further elucidate the costs associated with infection, or the genetic basis of resistance.

Here we present the isolation of KV from wild-collected *D. melanogaster* via passage in laboratory stocks and gradient centrifugation. We use this isolate to characterise the fundamental phenotypic impacts of infection on host longevity and fecundity. We then use the *Drosophila* Genetic Reference Panel (DGRP) (Mackay *et al.*, 2012) to quantify and dissect genetic variation in immunity to KV infection in males and females, and use RNA sequencing analyses of an inbred line to quantify host and virus transcriptional response in both sexes. We find that KV causes higher rates of mortality following injection in males, but that males have lower viral titre, suggesting some female tolerance to infection. However, we also find that female movement is decreased following infection, and that infected females have significantly reduced late-life fecundity—highlighting the importance of considering infection phenotypes beyond longevity. We find a genetic correlation in longevity between KV-infected males and females, and a weak negative genetic correlation between mortality and KV titre in females, and we report host loci that have variants significantly associated with each trait. Finally, our expression analysis of infected individuals supports a dramatic cessation of oogenesis following infection, and significant differential regulation of serine proteases and certain immune genes. This work establishes KV as a new natural model for DNA virus infection in *D. melanogaster* and will enable further dissection of the insect antiviral immune response.

4.3 Materials and Methods

4.3.1 Isolation of Kallithea Virus

We identified KV-infected flies through a PCR screen for previously published *D. melanogaster* viruses in 80 previously untested wild-caught flies (see Webster et al, 2015 for primers and cycling conditions). We homogenised each fly in 0.1 mL of Ringer's solution, transferred half of the homogenate to Trizol for nucleic acid extraction, and performed RT PCR assays on the

resulting RNA for all *D. melanogaster* viruses reported by Webster *et al* (2015). We selected a KV-positive sample from Thika, Kenya (Collected by John Pool in 2009; subsequently stored at -80C), removed debris from the remaining fly homogenate by centrifugation for 10 minutes at 1000 × g, and microinjected 50 nL of the supernatant into *Dicer-2*^{L811fsX} flies, which lack a robust antiviral immune response (Lee et al, 2004). After one week, we homogenised 100 KV-injected *Dicer-2* ^{L811fsX} flies in 10 uL Ringer's solution per fly, cleared the solution by centrifugation as above, and re-injected this homogenate into *Dicer-2*^{L811fsX} flies. This process was then repeated twice more with the aim of increasing viral titres. In the final round of serial passage, we injected 2000 *Dicer-2* ^{L811fsX} flies, which were homogenised in 5 mL 10 mM Tris-HCl. We cleared the homogenate by centrifuging at 1000 × g for 10 minutes, filtering through cheese cloth, centrifuging twice more at 6000 × g for 10 minutes, and finally filtering through a Millex 0.45 µm polyvinylidene fluoride syringe filter. The resulting crude virus preparation was used as input for gradient ultracentrifugation.

We screened the crude preparation by RT-PCR for other published *Drosophila* virus sequences, and identified the presence of DAV, Nora virus, DCV, and La Jolla virus. To separate KV from these viruses, we used equilibrium buoyant density centrifugation in iodixanol ("OptiPrep", Sigma-Aldrich) as enveloped viruses are expected to have lower buoyant densities than most unenveloped viruses. Iodixanol is biologically inert, and gradient fractions can be used directly for downstream infection experiments (avoiding dialysis, which we found greatly reduces KV titres). We concentrated virus particles by centrifuging crude virus solution through a 1 mL 10% iodixanol layer onto a 2 mL 30% iodixanol cushion at 230,000 × g for 4 hours in a Beckman SW40 rotor. Virus particles were taken from the 30%-10% interphase, and layered onto a 40%-10% iodixanol step gradient, with 2% step changes, and centrifuged for 48 hours at 160,000 × g. We fractionated the gradient at 0.5 mL intervals, phenol-chloroform extracted total nucleic acid from aliquots of each fraction, and measured virus concentration by quantitative PCR (qPCR). We pooled all Kallithea-positive, RNA virus-negative fractions and calculated the infectious dose 50 (ID50) by injecting 3 vials of 10 flies with each of a series of 10-fold dilutions and performing qPCR after 5 days. We simultaneously performed the above isolation protocol with uninfected *Dicer-2*^{L811fsX} flies and extracted the equivalent fractions for use as an injection control solution (hereafter referred to as "gradient control").

4.3.2 Transmission electron microscopy

A droplet of viral suspension was allowed to settle on a Formvar/Carbon 200 mesh Copper grid for 10 minutes. We removed excess solution and applied a drop of 1% aqueous uranyl acetate for 1 minute before removing the excess by touching the grid edge with filter paper. The grids were then air dried. Samples were viewed using a JEOL JEM-1400 Plus transmission electron microscope, and representative images were collected on a GATAN OneView camera.

4.3.3 Measurement and analysis of viral titre

Flies were reared on a standard cornmeal diet until infection, after which they were transferred to a solid sucrose-agar medium. We infected flies by abdominal injection of 50 nL of 10^5 ID50 KV using a Nanoject II (Drummond Scientific), and these flies were then used to assay changes in viral titre, mortality, fecundity, or daily movement. To test whether the change in viral titre over time was influenced by sex or the presence of *Wolbachia* endosymbionts, we injected 25 vials of 10 male or female *Oregon R* flies with KV, with or without *Wolbachia* (totalling 1000 flies). We phenol-chloroform extracted total nucleic acid at 5 time-points: directly after injection and 3, 5, 10, and 15 days post-infection. We used qPCR to measure viral titre relative to copies of the fly genome with the following (PCR primers: kallithea_126072F CATCAATATCGCGCCATGCC, kallithea_126177R GACCGAGTTAGCGTCAATGC, rpl32_465F CTAAGCTGTCGGTGAGTGCC, rpl32_571R: TGTTGTCGATACCCTTGGGC). We analysed the log-transformed relative expression levels of Kallithea virus as a Gaussian response variable in a linear mixed model using the Bayesian generalised mixed modelling R package MCMCglmm (V2.24) (Hadfield, 2010). R code and raw data used to fit all models in this paper is provided on figshare (doi: 10.6084/m9.figshare.c.3936037.v1).

The fixed effects portion of the model included an intercept term and coefficients for the number of days post-inoculation (DPI), sex, and DPI by sex interaction. We estimated random effects for each qPCR plate and assumed random effects and residuals were normally distributed. We initially fitted the model with *Wolbachia* infection status included as a fixed effect, however this term was not significant and was excluded from the final model.

We also attempted to infect flies with KV by feeding. We anesthetised flies in an agar vial and sprayed 50 uL of $5x10^3$ ID50 KV onto the flies and food. We then collected flies immediately

(for the zero time-point) and at 7 DPI and used the primers above to calculate relative KV titre.

4.3.4 Mortality following KV infection

We performed mortality assays to test the effect of KV infection on longevity, and to test whether this was affected by sex or *Wolbachia* infection status. We injected a total of 1200 *Oregon R* flies with control gradient or KV for each sex with or without *Wolbachia* (*Wolbachia* had previously been cleared by 3 generations of Ampicillin treatment and its absence was confirmed by PCR). We maintained flies for each treatment in 10 vials of 10 flies, and recorded mortality daily for three weeks. Mortality that occurred in the first day after infection was assumed to be due to the injection procedure and excluded from further analysis. We analysed mortality using an event-analysis framework as a generalised linear mixed model using MCMCglmm, with per-day mortality in each vial as a binomial response variable. We included fixed effects for DPI, DPI^2 (used to capture nonlinear mortality curves), KV infection status, the two-way interaction between DPI and KV infection status, the two-way interaction between DPI and sex, and the three-way interaction between DPI, KV infection status, and sex. We fitted vial as a random effect to account for non-independence among flies within vials, assuming these follow a normal distribution. As in the model for viral titre, we found no evidence for differences associated with *Wolbachia* infection, and *Wolbachia* terms were excluded from the final model. The higher rate of male mortality we observed was also confirmed in a second independent experiment using an outbred population derived from the Drosophila Genetic Reference Panel (DGRP; see below).

4.3.5 Fecundity following KV infection

We measured fecundity during early (1 and 2 DPI) and late (7 and 8 DPI) Kallithea virus infection. Virgin female flies from an outbred population derived from the DGRP (Mackay et al, 2012; created from 113 DGRP lines and maintained at a low larval density with non-overlapping generations) were injected with either KV, or with chloroform-inactivated KV as a control, and individually transferred to standard cornmeal vials. The following day we introduced a single male fly into the vial with the virgin female. We transferred the pair to new vials each day and recorded the number of eggs laid. Per-day fecundity was analysed in MCMCglmm as a Poisson response variable using a hurdle model, which models the probability of zeroes in the data and the Poisson process as separate variables. We included

fixed effects associated with KV infection status, infection stage (early or late), the interaction between KV infection and infection stage, and random effects associated with each fly pair (vial).

We analysed ovary morphology to examine whether changes in fecundity were detectable in ovaries. Flies were injected with either control virus solution or KV and kept on solid sucrose-agar medium vials. After 8 DPI, flies were transferred to vials with standard cornmeal medium supplemented with yeast. Two days later, we dissected ovaries in phosphate-buffered saline solution, fixed ovaries in 4% paraformaldehyde, and stained nuclei with DAPI. Ovaries were analysed under a Leica fluorescence microscope, and we recorded whether each ovariole within an ovary included egg chambers past stage 8 (i.e. had begun vitellogenesis), and whether any egg chambers within an ovariole exhibited apoptotic nurse cells. The probability of an ovariole containing a post-vitellogenic egg chamber was analysed using a logistic regression in MCMCglmm, with KV infection status as a fixed effect and the ovary from which the ovariole derived as a random effect. We analysed whether apoptotic nurse cells are associated with KV virus-infected ovary in the same way.

4.3.6 Daily movement following KV infection

We used a *Drosophila* Activity Monitor (DAM, TriKinetics; Pfeiffenberger et al, 2010) to measure per-day total movement of individual flies (Pfeiffenberger *et al.*, 2010). The DAM is composed of multiple hubs, each with 32 tubes containing a single fly, and movement is recorded on each occasion the fly breaks a light beam. We injected 96 female flies from an outbred DGRP population with either chloroform-inactivated KV or KV, randomly assigned these flies within and across 3 hubs, and measured total movement for one week. Movement was binned for each day and this per-day total movement was analysed in a linear mixed model as a Poisson response variable using MCMCglmm. We completely excluded flies that failed to move for a whole day or longer, assuming them to be dead. As before, we included fixed effects associated with DPI, KV infection status, and the interaction between KV and DPI. We included random effects associated with each fly (repeated measures) and each of the DAM hubs, and assumed each of these take values from a normal distribution.

4.3.7 Quantitative genetic analysis

The DGRP is a collection of highly inbred fly lines derived from a *D. melanogaster* population collected in Raleigh, North Carolina (Mackay *et al.*, 2012), and is widely used to estimate and

dissect genetic variation in complex traits in *Drosophila*. We measured KV titre in females and mortality following KV infection in both sexes for 125 DGRP lines, and estimated genetic (line) variances and covariances among these traits. To measure viral titre in the DGRP, we infected 5 vials of 10 flies for each line across 5 days, with a vial from each line being represented each day. After 8 DPI, living flies were killed and homogenised in Trizol for nucleic acid extraction and qPCR. To measure mortality following KV infection in the DGRP, we injected 3 vials of 10 flies of each sex and recorded mortality on alternate days until half the flies in the vial were dead (i.e. median survival time). Flies were transferred to fresh agar vials every 10 days. Mortality occurring in the first 3 DPI was assumed to be caused by the injection procedure and was removed from the analysis.

We fitted a multi-response linear mixed model in MCMCglmm to estimate heritability and genetic covariances among lines

$$y_{iklpqr}^{trait} = \beta_{p}^{trait} + \beta_{pk}^{trait:sex} + \mu_{pq}^{trait:date} + \mu_{l}^{plate} + \mu_{pkr}^{trait:sex:line} + \varepsilon_{iklpqr}^{trait:sex} \quad [1]$$

where y_{iklpqr}^{trait} is the log-transformed relative viral titre or the duration until median mortality (LT50). We only estimated sex-specific fixed effects ($\beta^{trait:sex}$) for LT50, because we did not measure titre in both sexes. The first part of the random effects model accounts for block effects due to date of injection ($\mu^{trait:date}$) and qPCR plate (μ^{plate}). We assumed a 2x2 identity matrix as the covariance structure for $\mu^{trait:date}$, with effects associated with each trait from independent normal distributions. Effects for the l^{th} plate were assumed to be normally distributed. The second part of the random effects model ($\mu^{trait:sex:line}$) estimates the variance in each trait across lines and was allowed to vary by sex. We estimated all variance-covariance components of the 3x3 G matrix associated with $\mu^{trait:sex:line}$. Finally, we fitted separate error variances for each trait in each sex ($\varepsilon^{trait:sex}$), where residuals were associated with independent normal distributions.

The diagonal elements of the $\mu^{trait:sex:line}$ covariance matrix represent posterior distributions of genetic variances for viral titre in females, LT50 in females, and LT50 in males (V_G^{titre}, $V_G^{Fmortality}$, $V_G^{Mmortality}$). We calculated broad-sense heritability (i.e. line effects) for each trait as $H^2 = \frac{V_G}{V_G+V_R}$, where V_R is the residual variance associated with each trait, estimated in the model as $\varepsilon^{trait:sex}$. However, heritabilities cannot readily be compared because of their dependence on the residual variance, which can be vastly different for different phenotypes (Houle, 1992). Therefore, we also calculated the coefficient of genetic

variation (CV_G) as $CV_G = \frac{100*\sqrt{V_G}}{\mu}$, where V_G is standardised by the phenotypic mean (μ) and is more appropriate for comparisons across phenotypes. All confidence intervals reported are 95% highest posterior density intervals.

4.3.8 Genome-wide association studies

We used measurements of viral titre and mortality following KV infection in the DGRP lines to perform a series of genome-wide association studies (GWAS). Although our power to detect small-effect genetic variants with only 125 lines is very low, past studies have demonstrated genetic variation in natural viral resistance in *Drosophila* is often dominated by few large effect variants (Contamine, Petitjean and Ashburner, 1989; Magwire *et al.*, 2011, 2012; Cogni *et al.*, 2016) but see (King and Long, 2017). We performed a GWAS on each phenotype separately by fitting an individual linear model for each variant in the genome using the full data. For the titre GWAS, we included focal SNP, qPCR plate, and date of injection as linear predictors. For the mortality GWAS, we included focal SNP, sex, and a sex-by-SNP interaction as linear predictors. Models were fitted using the base R linear model function 'lm()'. We tested the significance of the SNP and SNP-by-sex predictors with a t-test, and we obtained significance thresholds for each GWAS by permuting genotypes across phenotypes 1000 times and recording the lowest *p*-value for each pseudo-dataset.

4.3.9 Confirmation of GWAS hits

We chose 19 genes identified near significant GWAS hits to further test their involvement in KV infection. For each gene, we crossed a transgenic line containing a homologous foldback hairpin under the control of the Upstream Activating Sequence (UAS) to two GAL4 lines: *w*; P{UAS-3xFLAG.dCas9.VPR}attP40, P{tubP-GAL80^{ts}}10; P{tubP-GAL4}LL7/TM6B, Tb^{1}* (Bloomington line #67065; hereafter referred to as *tub-GAL4*) and *w*; P{GawB}Myo31DF^{NP0001}/CyO; P{UAS-3xFLAG.dCas9.VPR}attP2, P{tubP-GAL80^{ts}}2* (Bloomington line #67067; hereafter referred to as *myo31DF-GAL4*). These lines drive GAL4 expression in the entire fly and in the gut, respectively, and contain a temperature-sensitive Gal80, which is able to inhibit GAL4 at the permissive temperature (18 degrees). We used both broadly expressed and gut-specific GAL4 lines during systemic infection because previously described nudiviruses infect the insect midgut (Huger, 1966; Unckless, 2011), and because studies have shown natural and artificial infection routes can lead to similar tropism (for Drosophila C virus) and pathology (Oryctes rhinoceros nudivirus) (Huger, 1972; Zelazny,

1973a; Ferreira *et al.*, 2014).

RNAi lines included the following genes (BDSC numbers): *Pkcdelta* (28355), *btd* (29453), dos (31766), *tll* (34329), *Atg10* (40859), *Dgk* (41944), *Cip4* (53321), *hppy* (53884), *LpR2* (54461), *CG5002* (55359), *sev* (55866), *eya* (57314), *Gprk2* (57316), *Sox21b* (60120), *CG11570* (65014), *ATPCL* (65175), *Pdcd4* (66341), *CG7248* (67231), and *yin* (67334). As a control, we crossed the genetic background of the RNAi lines (Bloomington line #36304) to the two GAL4 lines. All crosses were made at 18 degrees. After eclosion, offspring were transferred to agar vials (10 flies per vial) at the non-permissive temperature (29 degrees) for two days to facilitate silencing of candidate genes, then injected with KV. We measured titre at 5 DPI for 5 vials of each KV-infected genotype for each GAL4 driver. We used a linear mixed model to analyse log-transformed viral titre in each knockdown relative to the genetic background controls, with GAL4 driver as a fixed effect, gene knockdown as a random effect, and with separate error variances for each GAL4 driver. If the random effect associated with a candidate gene was significantly different from zero, we concluded this gene played a role in determining the outcome of infection by KV. The specification of gene as a 'random effect' allows comparison of each knockdown to all other knockdowns, accounting for any possible overall effect of overexpressing a dsRNA hairpin. As a proof of principle, we confirmed knock-down of the largest-effect gene (*Cip4*) using the DRSC FlyPrimerBank qPCR primers *Cip4_PP33370F* (ATTGCGGGAGTGACGCTTC) and *Cip4_PP33370R* (CTGTGTGGTGAGGTTCTGCTG). We did not assess knockdown efficiency for the other crosses, and any negative findings should be treated with caution.

4.3.10 Sample preparation for RNA-sequencing

We next aimed to characterise the host expression response to KV infection, and whether this differed between males and females. We injected 6 vials of 10 flies for each sex with either the control gradient solution or with KV. After 3 DPI, we homogenised flies in Trizol, extracted total nucleic acid, and enriched the sample for mRNA through DNAse treatment and poly-A selection. We used the NEB Next Ultra Directional RNA Library Prep Kit to make strand-specific paired-end libraries for each sample, following manufacturer's instructions. Libraries were pooled and sequenced by Edinburgh Genomics (Edinburgh, UK) using three lanes of an Illumina HiSeq 4000 platform with strand-specific 75 nucleotide paired end reads. We subsequently identified a low level of Drosophila A Virus (DAV) contamination in both KV treated and untreated flies, reflecting the widespread occurrence of this virus in fly stocks

and cell cultures. All reads have been submitted to the European Nucleotide Archive under project accession ERP023609.

We used SPAdes genome assembler (v3.11.1) (Bankevich *et al.*, 2012) to assemble the KV genome from RNA-sequencing reads, using the previously published genome (NC_033829.1) as an 'untrusted contig' (File S1).

4.3.11 Differential expression analysis

We removed known sequence contaminants (primer and adapter sequences) from the paired end reads with cutadapt (V1.8.1) (Martin, 2011) and mapped remaining reads to the *D. melanogaster* genome (FlyBase release r6.15) and to all known *Drosophila* virus genomes using STAR (V2.5.3a) (Dobin *et al.*, 2013), with a maximum intron size of 100 KB, but otherwise default settings. We counted the number of reads mapping to each gene using the featurecounts command in the subread package (V1.5.2) (Liao, Smyth and Shi, 2013) and used these raw count data as input to DESeq2 (V1.16.0) (Love, Huber and Anders, 2014) for differential expression analysis. DESeq2 fits a generalised linear model for each gene, where read counts are modelled as a negative binomially distributed variable (Anders and Huber, 2010; Love, Huber and Anders, 2014) and includes a sample-specific size factor and a dispersion parameter that depends on the shared variance of read counts for genes expressed at similar levels (Anders and Huber, 2010; Love, Huber and Anders, 2014). Our design matrix included sex, KV infection status, and the interaction between the two, allowing us to test for expression changes following KV infection and how these changes differ between the sexes. To account for the unintended presence of DAV, and differences in the level of DAV within and between the treatments, we also include the relative titre of DAV as a continuous predictor. Using this model, we calculated $\log_2$ fold changes in DESeq2, and tested for significance using Wald tests. We used the 'plotPCA' function implemented in DESeq2 to perform principal component analysis of the rlog-transformed read count data (Love, Huber and Anders, 2014).

4.3.12 GO term and network analysis

We performed five independent gene ontology (GO) term enrichment analysis, using: (1) genes with significant SNPs in the GWAS for titre; (2) genes with significant SNPs in the GWAS for mortality, (3) genes upregulated in either sex ($p < 0.001$); (4) genes downregulated in either sex ($p < 0.001$); and (5), genes significantly different between males and females ($p <$

0.05). For each of these gene lists, we tested for GO term enrichment using the 'goseq' R package (V1.26.0) (Young *et al.*, 2010), which accounts for the difference in power for detecting differential expression caused by gene length, and tests for significant over-representation of genes in a GO term.

We performed a network analysis on genes identified in GWAS or RNA-sequencing studies at a liberal significance threshold ($p < 0.10$) to infer broadly acting pathways involved in KV infection that may have been overlooked in individual gene analyses. We used the PPI-spider tool (Antonov *et al.*, 2009) available on the bioprofiling webserver (Antonov, 2011), which uses the IntAct database to find enriched subpathways within a provided gene list, allowing one gene absent from the provided list to mediate an interaction. Enriched pathways in the given gene list are then compared to random gene lists of the same length to assess significance.

4.5 Results and Discussion

4.5.1 Isolation of Kallithea Virus

We isolated Kallithea Virus (KV) by gradient centrifugation following 4 rounds of serial passage in flies. Many laboratory fly stocks and cell culture lines are persistently infected with RNA viruses (Brun and Plus, 1980; Webster *et al.*, 2015), and following serial passage we identified co-infections of DAV, Nora Virus, and Drosophila C Virus (DCV) by PCR. The high prevalence of these viruses in laboratory stocks presents a substantial hurdle in the isolation of new *Drosophila* viruses, requiring the separation of the new viruses of interest. Although this can be relatively simple (e.g. separating enveloped from non-enveloped viruses), most of the recently identified *Drosophila* viruses (Webster *et al.*, 2015, 2016; Medd *et al.*, 2018) are from ssRNA virus families with buoyant densities similar to common laboratory infections. To exclude these from our isolate, we concentrated KV using a 1.18 g/mL cushion, retaining KV at the interphase, but excluding most of the contaminating RNA viruses. Subsequent equilibrium density gradient centrifugation produced a KV band at 1.17 g/mL, and with some DAV contamination at approximately 1.20 g/mL (Figure 1A). Although nudiviruses have not previously been prepared using an iodixanol gradient, the equilibrium buoyant density was consistent with the lower buoyant densities of enveloped particles (Feng *et al.*, 2013) and similar to other enveloped dsDNA viruses (e.g. Herpesviruses: 1.15 g/mL). KV was estimated to be an approximately 650-fold higher concentration than DAV at 1.17 g/mL, and we were unable to identify intact DAV particles by electron microscopy (KV shown in Figure 1B). KV is

morphologically similar to Oryctes rhinoceros nudivirus (Alois M Huger, 2005), with an enveloped rod-shaped virion approximately 200 nm long and 50 nm wide.

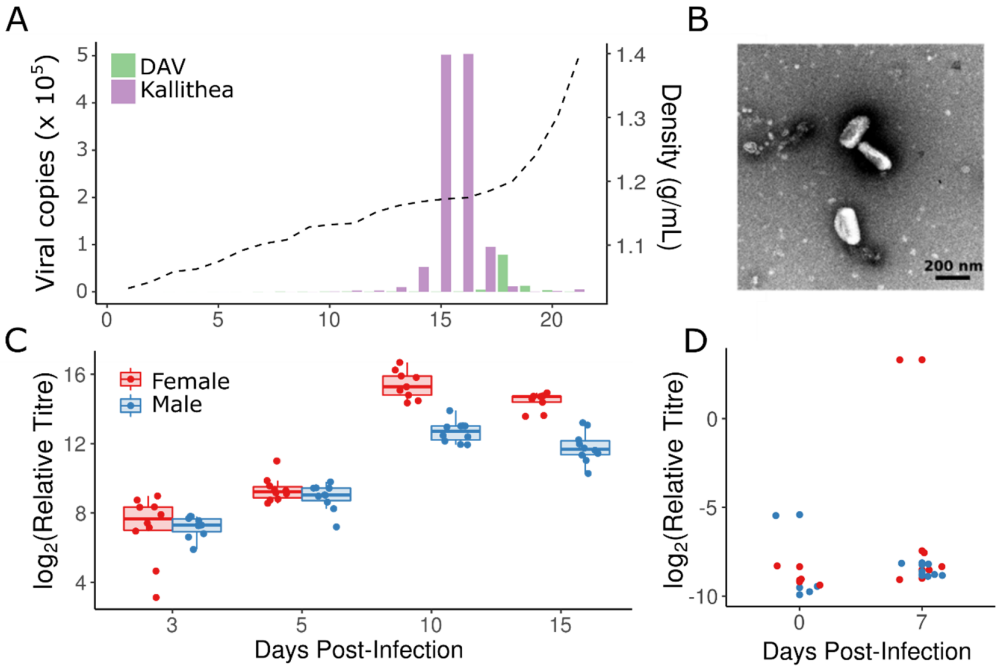

Figure 1: Isolation of KV and growth in flies

(A) Density gradient and virus titre: Kallithea virus (purple) was effectively separated from DAV (green) at 1.18 g/mL (dotted line) in fractions 15 and 16 of an iodixanol gradient. (B) Transmission electron micrograph of KV-positive fractions showed KV to be a rod-shaped enveloped particle, as has been described previously for other nudiviruses (Huger, 2005). We did not observe unenveloped KV particles, bacteria, or RNA viruses in the isolate. (C) Relative viral titres normalised by the number of fly genomic copies and virus levels at time zero in each sex. Each point represents a vial of 10 flies. Viral titres peaked at 10 days post-infection, and were generally higher in females (red) than males (blue) late in infection. (D) We were able to infect adult OreR flies orally by applying the viral isolate to Drosophila medium, although relative copy number of the virus was very low and infection was inefficient, with only 2 of 16 vials (each of 10 flies) having increased titre after one week, indicating an infectious rate lower bound of ~1% at $5x10^3$ ID50.

4.5.2 Kallithea virus growth in flies

We injected the KV isolate into *Drosophila Oregon R (OreR)* males and females, with and without *Wolbachia*, and measured viral titre at four time-points by qPCR. In females, KV increased approximately 45,000-fold by day 10, and then began to decrease, with reduced titre at 15 DPI (Figure 1C). In males, the KV growth pattern was altered, growing more slowly (or possibly peaking at an earlier un-sampled time point), resulting in a 7-fold lower titre than in females after 10-15 days, (nominal MCMC p-value derived from posterior samples, MCMCp= 0.002). *Wolbachia* did not affect virus growth rate in either sex across time points

(MCMCp = 0.552, Figure S1), reaffirming previous findings that *Wolbachia* do not offer the same protection against DNA viruses in *Drosophila* as they do against RNA viruses (Teixeira, Ferreira and Ashburner, 2008).

Nudiviruses have previously been reported to spread through sexual and faecal-oral transmission routes. The *Drosophila innubila* Nudivirus (DiNV), a close relative of KV, is thought to spread faecal-orally, so we tested whether KV can spread through infected food. We found that although oral transmission occurred in OreR flies at some point before 7 DPI, it was relatively inefficient (Figure 1D). However, the concentration of DiNV found in *D. innubila* faeces is broadly similar to our KV isolate after gradient centrifugation (Unckless, 2011) (Figure 1D), but the administered suspension had been diluted 50-fold and may consequently provide a lower dose than flies encounter naturally. To explore the potential for transovarial vertical transmission or gonad-specific infections following sexual transmission (as reported for Helicoverpa nudivirus 2) (Burand *et al.*, 2012), we also performed qPCR on dissected ovaries and the remaining carcasses at 3 DPI (Figure S2). We found that KV was highly enriched in the carcass relative to the ovaries. Although intra-abdominal injection could influence KV tissue-specificity, there were still substantial levels of KV in the ovaries, indicating there is not a complete barrier to infection. These results imply that KV is likely transmitted faecal-orally, as are closely related nudiviruses, but explicit tests for transovarial or sexual transmission are required.

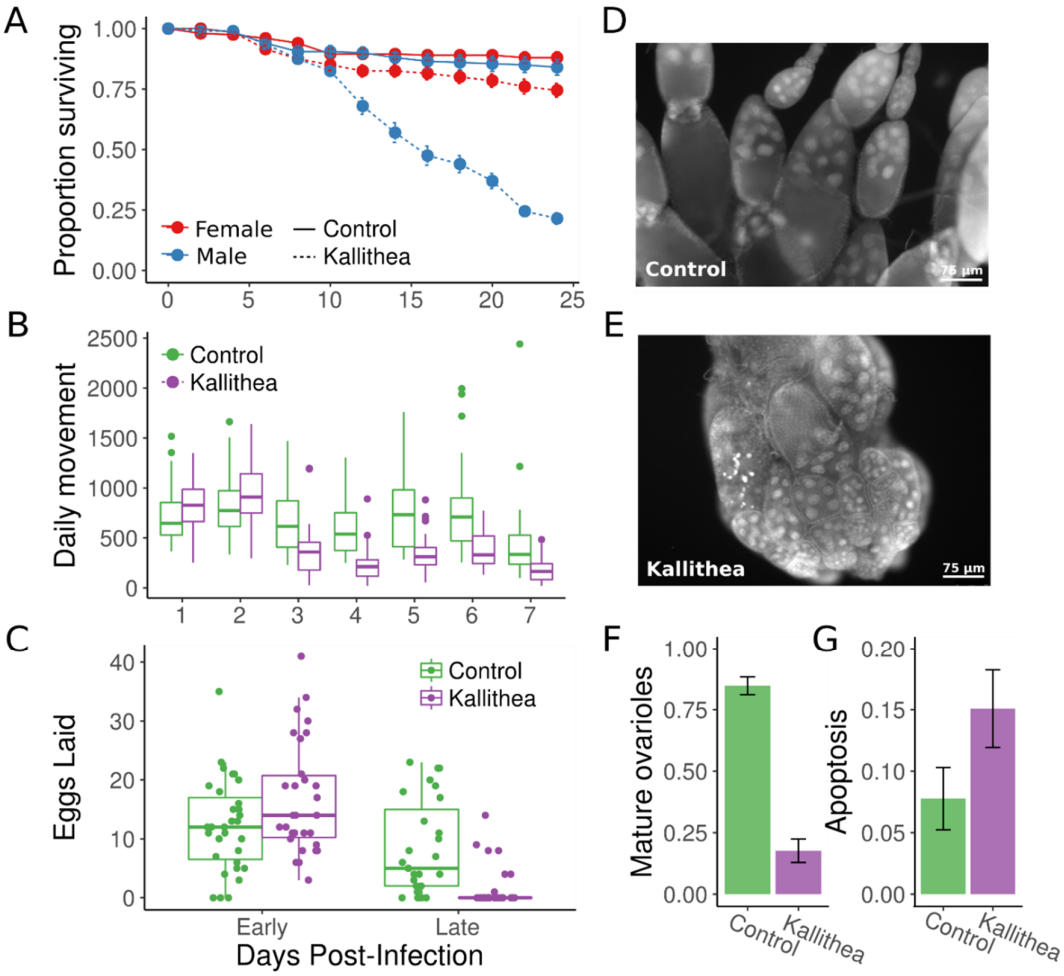

Figure 2: KV causes male-biased mortality, increased lethargy, and decreased fecundity

(A) Injection of KV virus into OreR flies led to sex-specific mortality. Infected females (red dotted line) experienced a small but significant increase in mortality, but males (blue dotted line) experienced a significantly larger rate of mortality after day 10. Flies injected with control gradient solution were unaffected (solid lines). Each point is the mean and standard error for the proportion of flies alive in each vial (10 vials of 10 flies). (B) Although females remained alive for longer, they were more lethargic. We assessed daily movement of flies from a DGRP outbred population injected with either chloroform-inactivated KV (green) or active KV (purple). KV-infected flies moved less from days 3-7 post-infection. (C) Females also displayed altered egg laying behaviour. Thirty pairs of flies from a DGRP outbred population were injected with inactive chloroform treated KV (green) or active KV (purple). KV-infected flies laid a slightly, but not significantly, higher number of eggs during early infection (1 and 2 DPI) but laid significantly fewer eggs in late infection (7 and 8 DPI). This reduction in egg laying is due to a shutdown of oogenesis before vitellogenesis (D, E), and ovaries from KV-infected flies house a lower proportion of ovarioles that include late-stage and mature egg chambers (F) and a higher proportion which contain apoptotic nurse cells (G). Ovaries were analysed 10 DPI, and error bars (F,G) show the standard error.

4.5.3 Sex-specific mortality, lethargy, and altered fecundity patterns following KV infection

Drosophila innubila infected with DiNV suffer fitness costs including increased mortality and decreased fecundity (Unckless, 2011). We investigated KV-induced mortality in *D. melanogaster* by injecting *OreR* males and females, with and without *Wolbachia*, with either control gradient solution or KV. We found that KV caused slightly, but significantly, increased mortality in females compared with controls (21% dead by day 20, vs. 11% in controls, MCMCp = 0.001), but caused a dramatically increased mortality in males compared to females (63% dead by day 20, vs. 14% in controls, sex:virus interaction MCMCp < 0.0001; Figure 2A). Therefore, males appear less tolerant of infection by KV, displaying increased mortality and a lower titre than females. We confirmed the KV-induced male death was not caused by DAV or other unknown small unenveloped RNA viruses present in our initial isolate, as chloroform treatment of the KV isolate eliminated treatment associated mortality (Figure S3). Male-specific costs of infection are widespread across animal hosts and their pathogens (Zuk, 2009), and reduced male tolerance has been found in flies infected with DCV (Gupta *et al.*, 2017). We found that *Wolbachia* infection had no detectable effect on KV-induced mortality in males or females, and thus does not affect tolerance (MCMCp = 0.20; Figure S1). This is consistent with previous studies showing that *Wolbachia* infection affects resistance and tolerance to RNA viruses but not a DNA virus (Teixeira, Ferreira and Ashburner, 2008).

We next tested whether female flies suffer sub-lethal fitness costs, by monitoring fly movement for a week following infection. KV-infected female flies showed similar movement patterns to chloroform-treated KV-injected flies for two days post-infection, but from three days post-infection moved significantly less (~70% reduction relative to controls; MCMCp < 0.001; Figure 2B). We conclude that females suffer from increased lethargy resulting from KV infection. In a natural setting, this could translate into fitness costs associated with increased predation, and reduced egg dispersal, mating, and foraging.

Finally, we tested whether KV infection resulted in decreased fecundity by monitoring the number of eggs laid by female flies derived from an outbred DGRP population for 8 days post-infection. We found that infected females exhibited markedly different egg laying patterns (MCMCp < 0.001; Figure 2), with KV-infected flies consistently laying fewer eggs between 7 and 8 days post-inoculation. This reduction in egg-laying during late infection could be due to a behavioural response or a cessation of oogenesis. To differentiate between these

possibilities, we dissected ovaries, and determined the proportion of ovarioles that contained mature egg chambers. We found that ovaries from KV-infected flies (10 DPI) halt oogenesis around stage 8 (MCMCp < 0.001), before vitellogenesis, and house an increased number of apoptotic egg chambers (MCMCp < 0.001) (Figure 2). This phenotype is similar to that seen upon starvation (Jouandin, Ghiglione and Noselli, 2014), and could be the manifestation of a trade-off to reroute resources to fighting infection, or of sickness-induced anorexia (Ayres and Schneider, 2009). Alternatively, this could be a direct consequence of viral infection, consistent with the gonadal atrophy reported for HzNV-2 (Burand *et al.*, 2012). Future studies should address whether this phenotype is a direct or indirect consequence of infection, and if the latter, whether it is orchestrated by the host or the virus.

Trait	Mean	V_G	V_P	H^2	CV_G
Titre	1.96 [1.31 - 2.63]	0.05 [0.03 - 0.09]	0.28 [0.24 - 0.32]	0.19 [0.1 - 0.29]	11.8 [8.4 - 15.2]
LT50 ♀	25.8 [22.6 - 29.0]	25.2 [17.4 - 33.8]	44 [35.7 - 52.6]	0.57 [0.47 - 0.67]	19.5 [16.2 - 22.6]
LT50 ♂	19.6 [16.4 - 22.7]	10.9 [5.9 - 16.5]	35.7 [30.1 - 41.7]	0.3 [0.18 - 0.42]	16.7 [12.7 - 21]

Table 1: Trait means, genetic variance (V_G), total phenotypic variance (V_P), heritability (H^2), and coefficient of genetic variation (CV_G) in titre and mortality following KV infection in the DGRP

4.5.4 Variation in titre and mortality following KV infection

The DGRP (Mackay *et al.*, 2012) have previously been used to dissect genetic variation underlying resistance and tolerance to bacterial, fungal, and viral pathogens (Magwire *et al.*, 2012; Bou Sleiman *et al.*, 2015; Howick and Lazzaro, 2017; Wang, Lu and St. Leger, 2017). We infected 125 DGRP lines with KV and estimated broad-sense heritabilities (H^2: the proportion of phenotypic variance attributable to genetic line) and coefficients of genetic variation (CV_G: a mean-standardised measure of genetic variation) in viral titre and LT50 values in females, and LT50 values in males (Table 1). Although female mortality following KV infection is not as pronounced as in males, we observed significantly increased mortality following KV infection in OreR females (MCMCp < 0.001, Figure 1), and reasoned that any genetic variation in this phenotype would be reflected in LT50 values. Our estimates of H^2 and CV_G fall within the range found for resistance to other pathogens in the DGRP, although direct comparison is difficult as studies are inconsistent in the statistics used to report genetic variation. H^2 in

survival following infection with an opportunistic bacterium or fungus was similar to our estimate for survival following KV infection (*Pseudomonas aeruginosa*: H^2 in males = 0.47, H^2 in females = 0.38; *Metarhizium anisopliae*: H^2 in males = 0.23, H^2 in females = 0.27), although comparing heritability can be easily confounded by differences in environmental (residual) variance (Houle, 1992; Wang, Lu and St. Leger, 2017). Genetic variation in resistance has also been measured in response to two non-native *D. melanogaster* viruses (Flock House Virus and *Drosophila affinis* Sigma Virus) and two native viruses (DCV and DmelSV) in females of the DGRP. Of these, the lowest heritabilities are those associated with resistance to non-native fly viruses (FHV: narrow sense heritability h^2 = 0.07, CV_G = 7; *D. affinis* sigma virus: h^2 = 0.13), and the highest are associated with native fly viruses (DCV: h^2 = 0.34, CV_G = 20; DmelSV: h^2 = 0.29). Although Magwire et al (2012) inferred h^2 as half V_G and accounted for the homozygosity of inbred lines when inferring CV_G, it is clear that V_G for resistance to KV is closer to the V_G for resistance to other native fly viruses than to non-native ones, at least for survival. It is also notable that CV_G estimates for survival are higher than estimates for titre, consistent with the observation that traits more closely related to fitness are expected to have higher CV_G values (Houle, 1992).

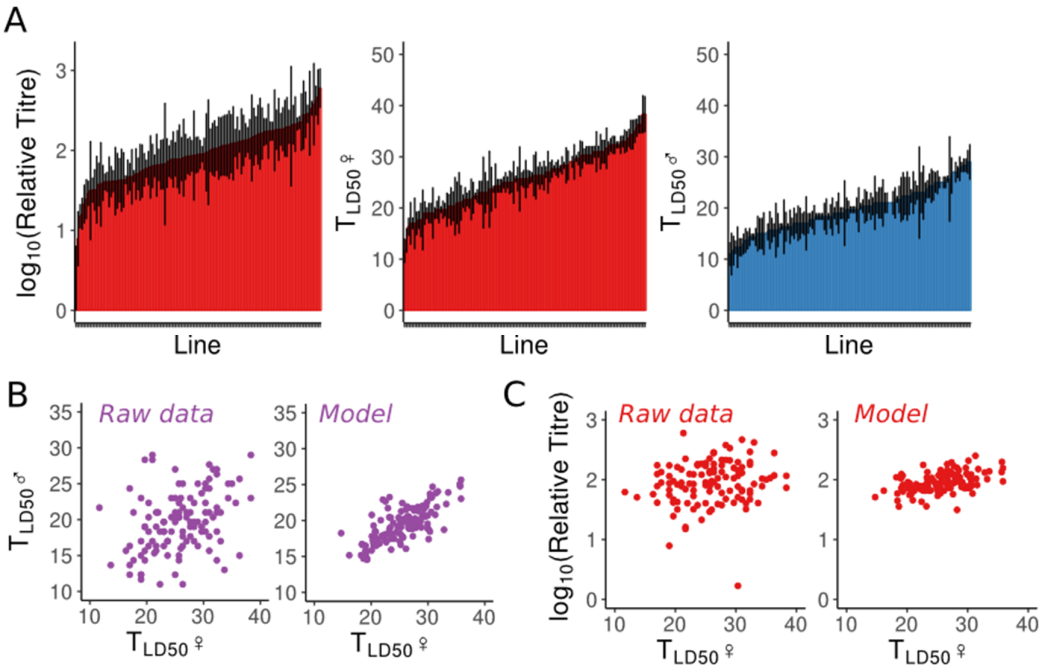

Figure 3: Genetic variation in resistance to KV

(A) We measured LT50 in both sexes, and titre in females, following KV injection in the DGRP. For titre, each bar represents the mean (and standard error) titre relative to fly genome copy-number, as assessed by qPCR for 5 vials of 10 flies for each of 125 DGRP lines. For LT50, each bar represents the mean time until half the flies (in a vial of 10) were dead, for three vials per line, per sex. (B, C) We used a multi-response linear mixed model to calculate genetic correlation between the traits. Shown are the raw data (left), and the estimated line effects (right) after accounting for any injection date and qPCR plate effects, and for the estimated variance among lines. Each point is a DGRP line measured for both phenotypes. We find a strong positive correlation between male and female LT50 values (B). We also observe a weak positive correlation between titre and LT50 (C).

We calculated genetic correlations between male and female mortality, and between viral titre and mortality in females (Figure 3). Note that we found no correlation between survival time following KV infection and published estimates of longevity in the absence of infection, nor to resistance to any other RNA viruses (Magwire *et al.*, 2012; Ivanov *et al.*, 2015). We found a strong positive correlation between males and females in median survival time following KV infection (0.57 [0.34-0.78]; MCMCp <0.001), such that lines in which infected males die quickly are also lines in which infected females die quickly, suggesting a shared genetic basis for early lethality following infection. We also surprisingly find a positive genetic correlation between viral titre and LT50 values (r = 0.32 [0.05-0.59], MCMCp = 0.017), such that fly lines that achieved higher titres on day 8 tended to live slightly longer. However, the effect size is small (a doubling of viral titre led to a half-day increase in median survival time) and the result is only marginally significant. The absence of a negative correlation is counter-

intuitive, and contrasts with infection of the DGRP with *Providencia rettgeri* and *Metarhizium anisopliae*, and infection across *Drosophila* species with DCV, where fly lines or species with higher parasite loads suffer increased mortality (Longdon, Hadfield, *et al.*, 2015; Howick and Lazzaro, 2017; Wang, Lu and St. Leger, 2017). This apparent decoupling of titre and mortality could result from inherent costs associated with the induction of an immune response, whereby flies that raise a more potent immune response keep KV at lower titres but induce greater tissue damage. Alternatively, KV-induced mortality in females may be a minor force relative to death caused by malnutrition on a sugar-only diet. If the reduced tolerance of males to KV was driven in part by malnutrition, as might be expected from other nudivirus infections that cause massive over-proliferation of midgut cells (Zelazny, 1973b, 1977; Alois M. Huger, 2005), then a positive genetic correlation between viral titre and survival of both sexes could result, if, for example, flies which fare worse on sugar diets are not as fit for viral replication.

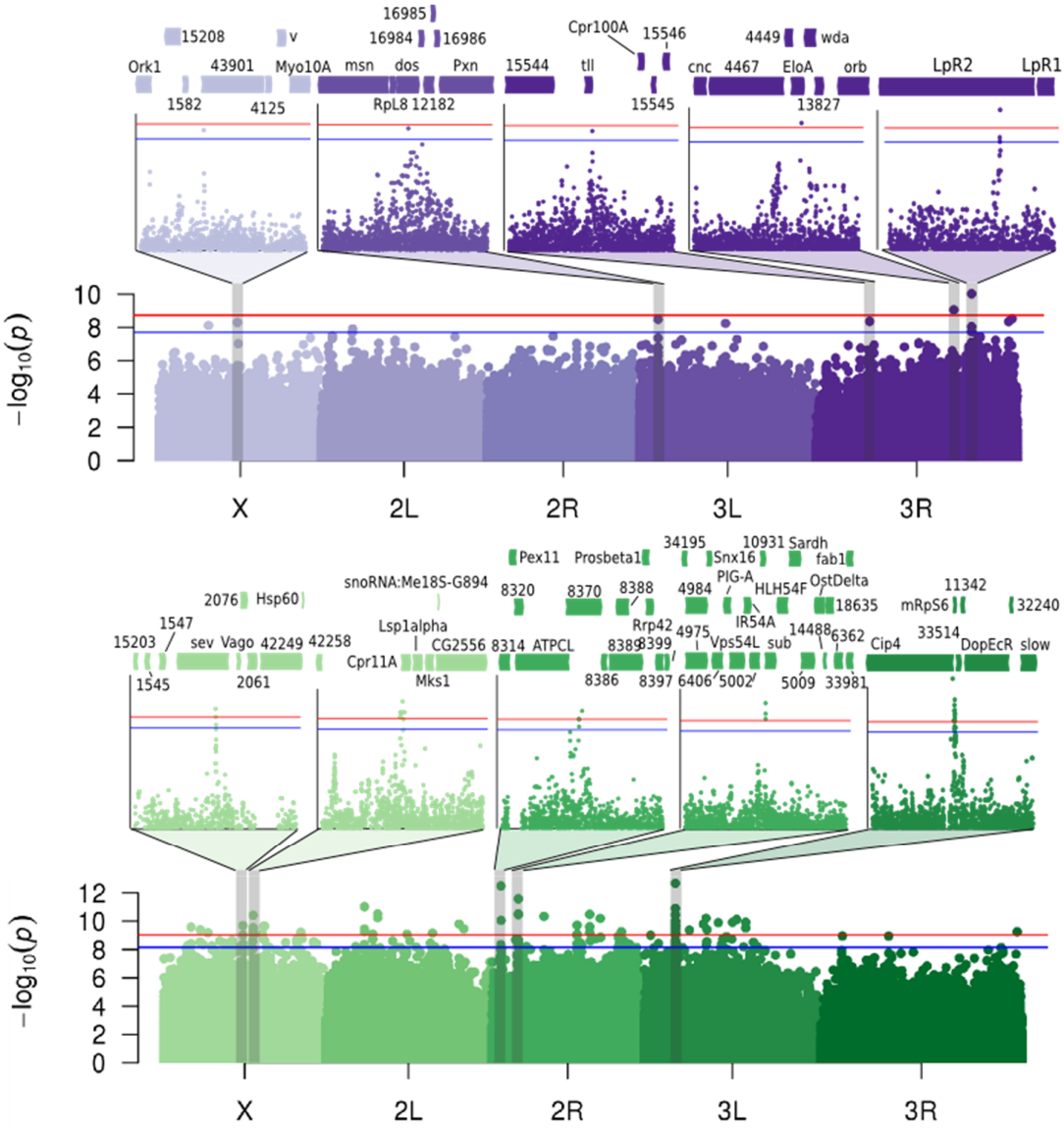

Figure 4: Genome-wide association of polymorphism in the DGRP with KV-induced titre and mortality

Manhattan plots showing the p-value for the effect of each polymorphism on viral titre (purple) and mortality (green). The top SNPs for each phenotype are shown in expanded inset panels, including surrounding genes. For clarity "CG" is omitted from gene identifiers. Horizontal lines show significance thresholds obtained through randomisation (p_{rand} = 0.05 in blue; p_{rand} = 0.01 in red).

4.5.5 Identification of candidate genes underlying host variation in KV titre

Using the phenotypes in the DGRP lines measured above, we performed a genome-wide association study to identify candidate genes underlying variation in titre, LT50, and differences in LT50 between the sexes. We found 10 SNPs (9 near genes) that were significantly associated with viral titre ($p_{rand} < 0.05$, based on 1000 random permutations of

phenotypes across lines; Figure 4). The SNP with the smallest p-value appeared in *Lipophorin receptor 2* (*LpR2*), which encodes a low-density lipoprotein receptor, previously found to be broadly required for flavivirus and rhabdovirus cell entry (Agnello *et al.*, 1999; Albecka *et al.*, 2012; Finkelshtein *et al.*, 2013).

We tested whether these candidate polymorphisms were enriched in any molecular, biological, or cellular processes using a GO enrichment analysis, and found the top hit to be the torso signalling pathway with 2 genes of 34 in the category ($p = 0.0004$), *tailless* and *daughter of sevenless* (*dos*). Torso signalling is upstream of extracellular-signal-regulated kinase (ERK) pathway activation in some tissues, and human orthologues of *dos* (GAB1/GAB2/GAB4) are cleaved by an enterovirus-encoded protease, thereby activating ERK signalling and promoting viral replication (Deng *et al.*, 2015, 2017). ERK signalling is also an important regulator of virus replication in the fly midgut, where it couples nutrient availability with antiviral activity (Xu *et al.*, 2013; Liu, Hodgson and Buchon, 2017). See Table S1 for a list of all nominally significant SNPs with associated locations, mutation types (e.g. intronic, synonymous coding, etc), nearby genes, p-values, effect sizes, and GO terms.

4.5.6 Identification of candidate genes underlying host variation in KV-induced mortality

We found 86 SNPs (65 near genes) that were significantly associated with LT50 following KV infection in the DGRP ($p_{rand} < 0.05$; Figure 4, Table S1), none of which were identified in the GWAS for viral titre. We performed a GO enrichment analysis, and found genes associated with these SNPs were enriched for hydrolase activity (top molecular function GO term, $p = 0.0004$), stem cell fate determination (top biological process GO term, $p = 0.002$), and in the plasma membrane (top cell component GO term, $p = 0.004$), among others (Table S2). Of these 86 SNPs, we found 34 (26 near genes) that were highly significant, and selected these for further analysis and confirmation ($p_{rand} < 0.01$; see Table S1 for all significant SNPs in). The polymorphism with the most confident association was located in *Cdc42-interacting protein 4* (*Cip4*), a gene involved in membrane remodelling and endocytosis (Leibfried *et al.*, 2008; Fricke *et al.*, 2009). This SNP is intronic in the majority of *Cip4* transcripts, but represents a nonsynonymous polymorphism segregating leucine and proline in the first exon of *Cip4-PB* and *Cip4-PD* isoforms, perhaps indicating spliceform-specific effects on KV-induced mortality. Of particular interest from the remaining 33 highly significant SNPs was a synonymous SNP in the receptor tyrosine kinase, *sevenless*, known to interact with *dos* (above), and seven genes

(*Dgk*, *Atg10, ATPCL, Hppy, Pkcdelta, Gprk2, Pdcd4*) previously implicated in viral pathogenesis or general immune processes. Of these, three (*Gprk4, hppy, Pkcdelta*) are involved in NF-κB signalling (Chuang *et al.*, 2011; Loegering and Lennartz, 2011; Valanne, Wang and Ramet, 2011). *ATPCL* was identified in an RNAi screen for factors regulating Chikungunya virus replication in humans (Karlas *et al.*, 2016) and is involved in the late replication complexes of Semliki Forest Virus (Varjak *et al.*, 2013). Finally, *Atg10* and *Pdcd4* are involved in autophagy and apoptosis, respectively, both broadly antiviral cellular functions known to have a role in antiviral immunity in *Drosophila* (Shelly *et al.*, 2009; Lamiable and Imler, 2014). We found no SNPs significantly associated with sex-specific KV-induced mortality (Figure S4).

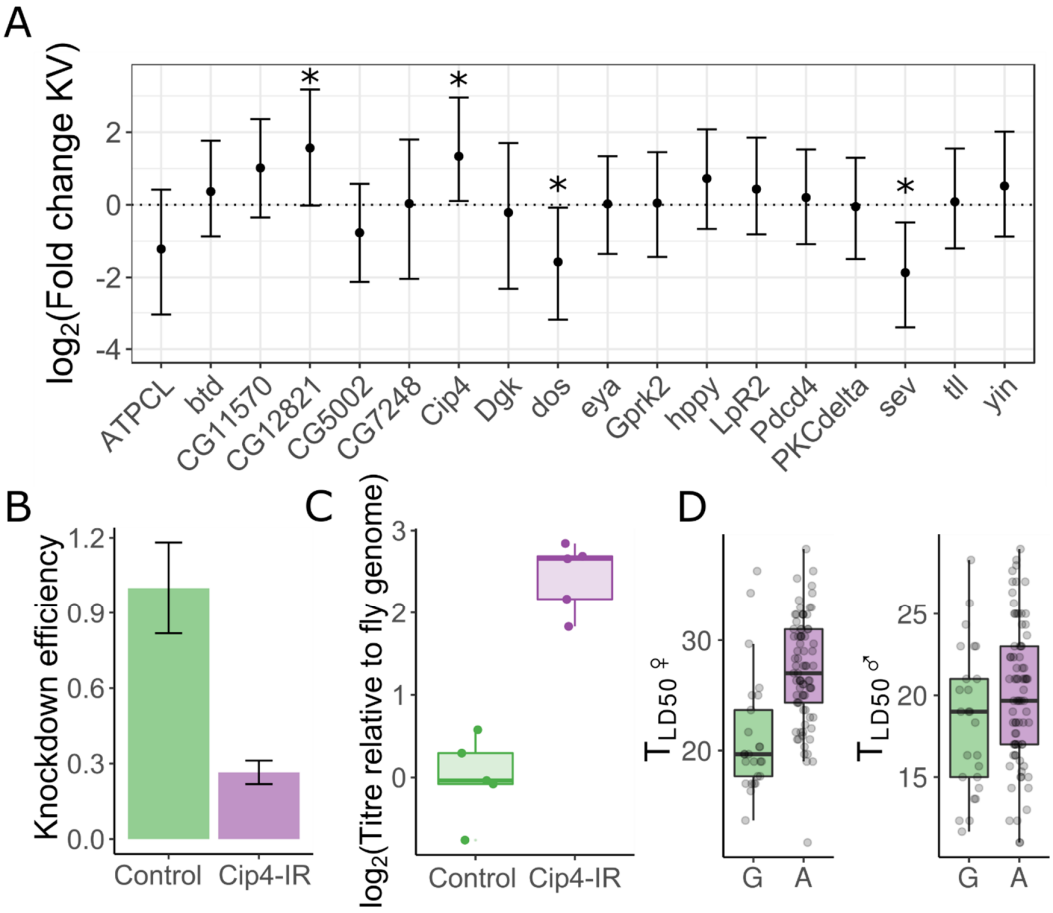

Figure 5: Confirmation of antiviral genes identified in GWAS

*KV titre was measured in flies expressing a foldback hairpin targeting 18 genes identified in the GWAS, using GAL4 lines that knock each down in either the whole fly or specifically in the gut. (A) The data were used to estimate random effects associated with each gene knock down, plotted with 95% highest posterior density intervals. (B) Knock-down of the most confident association in the GWAS, Cip4, caused reduced Cip4 RNA levels and (C) increased viral titre. (D) The associated variant (3L_4363810_SNP), was polymorphic (G/A), representing a nonsynonymous polymorphism in some splice variants, and survival following KV infection was significantly increased in fly lines with the "A" genotype, especially in females. Each point in comparison of survival in the two genotypes is a line mean. (*MCMCp < 0.05)*

4.5.7 Confirmation of GWAS hits

We chose 18 GWAS-candidate genes with available UAS-driven RNAi constructs to verify their involvement in KV infection. We found that knockdown of *Cip4* and *CG12821* caused significantly increased viral titre (2.6 and 2.9-fold change greater viral titre, respectively), and knockdown of *sev* and *dos* resulted in significantly decreased viral titre (0.27 and 0.33-fold reduction in viral titre, respectively), relative to other knockdown lines (Figure 5; Figure S5). We confirmed *tub-GAL4>Cip4^{IR}* flies had reduced (26% of wild-type) *Cip4* RNA levels and a concomitant increase in viral titre relative to the genetic background control (3.4-fold increase, 95% C.I. 1.3 – 9.6 fold) (Figure 5). This strongly suggests that *Cip4* is a KV restriction factor that likely segregates for functional polymorphism affecting survival following KV infection (Figure 5). It is known that baculovirus budded virions enter cells through clathrin-mediated endocytosis or micropinocytosis (Long *et al.*, 2006; Kataoka *et al.*, 2012), and gain their envelope at the cell membrane upon exit (Blissard and Rohrmann, 1990). Cip4 could therefore plausibly enact an antiviral effect by limiting KV cell entry or spread, perhaps through its known function in cell membrane remodelling and trafficking.

4.5.8 Differential expression following KV infection

Previous transcriptional profiling in response to RNA virus infection has shown upregulation of heatshock proteins, JAK-STAT, JNK, and Imd pathways (Dostert *et al.*, 2005; Kemp *et al.*, 2013; Zhu, Ding and Zhu, 2013; Merkling *et al.*, 2015). However, the *D. melanogaster* expression response to a DNA virus has not previously been investigated. We separately injected male and female *OreR* flies with control gradient solution or KV and extracted mRNA for sequencing 3 days post-infection. KV gene expression increased dramatically 3 days post-inoculation, consistent with our qPCR analysis of genome copy-number (Figure 1, Figure S10). Although not previously detectable by PCR, RNAseq read mapping identified a low level of DAV in both control and KV-infected flies, with an overall higher level in KV-infected flies. To account for this potentially confounding contaminant, we fitted the number of DAV-mapped reads as a covariate in the differential expression analysis, and used a stringent Benjamini–Hochberg adjusted significance threshold of $p < 0.001$ to infer nominal significance. We found 54 genes upregulated and 79 genes downregulated in response to KV in either males or females (Figure 6; Table S3). There was no enrichment for GWAS hits among the KV-

responsive genes (Figure S6). Principal components analysis on depth-normalised read counts separated males and females along PC1 and partially separated KV-infected and control-injected libraries along PC3 (Figure S7). GO term analysis identified 'defense response to virus' ($p = 3.1x10^{-4}$), 'serine peptidase activity' ($p = 1.2x10^{-7}$, identified in part due to downregulation of Jonah family serine proteases), and 'chorion' ($p < 1x10^{-8}$) as the most highly enriched biological process, molecular function, and cellular component, respectively (Figure 6; Table S4). Subsequent network analysis identified a large pathway of interactions enriched for genes either differentially expressed or associated with variation in KV infection, including known defense response genes (Figure S8).

There are few described induced antiviral immune effectors in *Drosophila* (Lamiable and Imler, 2014). In line with this we observe 57% of differentially expressed genes have not yet been named (i.e. "CG" genes), significantly greater than the genome-wide rate of 41% ($p = 3x10^{-4}$), and the most highly induced genes have not been implicated in viral pathogenesis. The cytochrome P450 family gene *Cyp304a1* was most highly upregulated, concomitant with the upregulation of four other genes in this family (*Cyp309a1, Cyp309a2, Cyp4p3*, and *Cyp6a20*). The next most highly induced genes include the hemocyanin *Larval serum protein 2*, the cytidine deaminase *CG8353*, four genes without functional annotation or recognisable domains (*CG33926*, *CG31955*, *CG32368*, *CG13641*), and an additional six genes without functional annotation (*CG43064, CG42825, Gagr, CG10211, CG17264*, and *CG17224* –the last two of which are adjacent on chromosome arm 2L). We also note the striking but variable upregulation of 11 of the 24 Tweedle genes (Figure S9) in some (but not all) of the infected samples. These are secreted, insect-specific cuticle proteins that regulate body shape (Guan *et al.*, 2006), and are also upregulated in response to Sindbis virus infection in cell culture (Graveley *et al.*, 2011), perhaps suggesting a general role in viral pathogenesis.

Genes with known involvement in viral infection were also found to be induced following KV infection. The RNAi effector *AGO2* was upregulated, consistent with the previous results that DNA viruses are a target of the RNAi pathway (Bronkhorst *et al.*, 2012; Kemp *et al.*, 2013; Alfred W Bronkhorst *et al.*, 2014). *Vago*, an antiviral factor downstream of Dicer-2 (Deddouche *et al.*, 2008), was upregulated and was also adjacent to a SNP found in the mortality GWAS (*dos*; Figure 4, Figure 7), as were *pastrel* and *ref(2)P*, identified in previous genome wide association analyses for resistance to DCV and DMelSV, respectively. Finally, we found that KV induced expression of *CG1667*, the *Drosophila* homologue of STING. The

vertebrate cGAS-STING pathway is involved in cytosolic DNA sensing and activation of immune factors in response to DNA virus infection (Chen, Sun and Chen, 2016). This upregulation of *CG1667* may suggest that this is another pathway conserved between *Drosophila* and vertebrates.

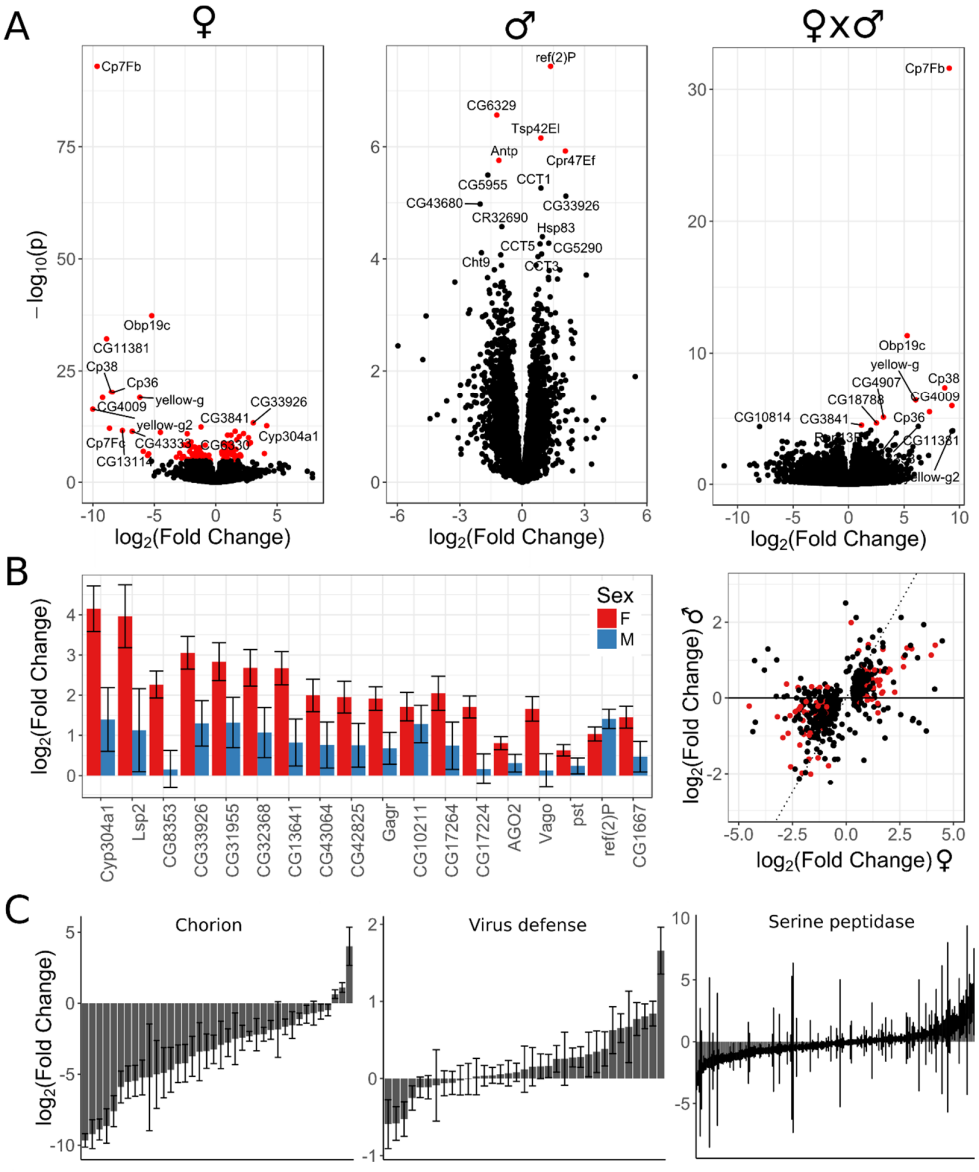

Figure 6: KV induces differential regulation of chorion, virus defense, and serine endopeptidase genes

(A) Volcano plots showing fold changes and p-values from Wald tests for differential expression of D. melanogaster genes 3 days following KV infection for OreR females (left), OreR males (center), those different between the sexes (right), with DAV read count fit as a covariate and nominal significance threshold of $p < 0.001$. In each panel, the genes with the smallest p-values are labelled. All of the genes that were significantly differentially regulated between the sexes are highly significant in females. (B) Highly induced genes are mostly functionally unannotated, but include some with known roles in viral pathogenesis. (C) The male response is correlated with females, but muted, with few genes identified as significantly differentially expressed in males. Genes with weak evidence of

differential expression in either sex ($p < 0.05$) are plotted, where the dotted line represents a perfect correlation, and red points are genes identified as significantly differentially expressed. (D) The top GO enrichment terms for each GO class (Molecular Function, Biological Process, Cellular Component) were genes involved the chorion, virus defense, and serine peptidase activity. For each plot, estimated fold changes and their associated standard errors are plotted for every gene matching the GO term, regardless of the significance of the Wald test. Generally, chorion genes were downregulated, virus defense genes were upregulated, and serine peptidases were downregulated.

As we had observed male and female differences in KV-induced mortality and titre (Figure 1, Figure 2), we tested for sex-specific transcriptional regulation in response to KV infection. We found that females and males had similar patterns of differential expression following KV infection (spearman rank correlation coefficient, $\rho = 0.57$, $p = 2.2 \times 10^{-16}$), although the male response was often less potent (Figure 6). Nine genes were significantly differentially expressed ($p < 0.05$) between the sexes specifically in response to KV (Table S3), and these were all downregulated in females and highly enriched for genes associated with the chorion (Figure 6, Table S4). Strikingly, all but three genes classified with the GO term 'chorion' were downregulated in females (Figure 6), consistent with the observed reduction in mature ovarioles and eggs during late infection, and implying a substantial reorganization of oogenesis (Figure 2). We did not identify any previously described immune genes with significant sex-specific regulation during KV infection.

4.5.9 Conclusions

We have isolated Kallithea virus, a dsDNA nudivirus that naturally infects *D. melanogaster*, and find it to be experimentally tractable. KV infection leads to reduced fertility and movement in females, highlighting the importance of measuring fitness associated traits besides longevity. Although males suffered greater mortality than females, they achieved lower titres, consistent with increased resistance and/or reduced tolerance in males. Similar to RNA viruses, we identified moderate host genetic variation in resistance to KV infection, however, we found that the underlying genetic architecture of this variation is unlike previously studied RNA viruses of *D. melanogaster*, in which a high proportion of genetic variation was apparently determined by a small number of loci. This could reflect a difference in the co-evolutionary dynamics between *D. melanogaster* and KV, versus other RNA viruses such as DCV and DMelSV. The *D. melanogaster* transcriptional response to KV included genes with known involvement in viral pathogenesis, but also genes that could represent infection responses distinctive to DNA viruses or KV, including downregulation of chorion genes. Upregulation of widely conserved immune factors, such as *STING*, represent promising

candidates involved in fly antiviral immunity, and demonstrate the continued utility of the *Drosophila* system for understanding host-virus interactions.

Chapter 5: Induction and suppression of NF-κB signalling by a DNA virus of Drosophila

The mass spectrometry was performed by Joep Joosten, who also provided an initial draft of the associated methods section. I am the sole author of the remaining text, with comments on earlier drafts from Darren Obbard.

5.1 Abstract

Interactions between the insect immune system and RNA viruses have been best studied in *Drosophila*, where RNA interference, NF-κB and JAK-STAT pathways underlie antiviral immunity. In turn, insect viruses have convergently evolved suppressors of RNA interference, which act by diverse mechanisms to allow viral replication. However, interactions between the insect immune system and DNA viruses have received less attention, primarily because few *Drosophila*-infecting DNA virus isolates are available. Here, we use a recently-isolated DNA virus of *Drosophila* melanogaster, Kallithea virus, to probe known anti-RNA viral immune responses and virus evasion tactics in the context of DNA virus infection. We find that flies mutant for RNA interference and Immune deficiency (Imd), but not Toll, pathways are more susceptible to Kallithea virus infection. However, we identify Kallithea virus-encoded gp83 as a potent inhibitor of Toll signalling, strongly suggesting Toll could mediate antiviral responses during Kallithea virus infection, but that it is suppressed by the virus. Further, we found Kallithea virus-encoded gp83 inhibits Toll signalling either through NF-κB transcription factor regulation, or transcriptionally. Together, these results provide a broad description of known antiviral pathways in the context of DNA virus infection, and extend the known diversity of insect virus-encoded immune inhibitors.

5.2 Introduction

Innate antiviral immunity in insects has been best studied in response to RNA virus infections of *Drosophila melanogaster*. Antiviral immune mechanisms that target RNA viruses include RNA-mediated defenses such as RNA interference (RNAi) and decay, cellular defences such as apoptosis, phagocytosis, and autophagy, and transcriptional responses. The latter are primarily mediated by Janus kinase/signal transducers and activators of transcription (JAK-STAT)

and Nuclear factor κB (NF-κB) pathways (reviewed in Merkling and van Rij, 2013; Bronkhorst and van Rij, 2014; Lamiable and Imler, 2014; Xu and Cherry, 2014; Palmer, Varghese and van Rij, 2018).

The *Drosophila* response to DNA viruses is less well studied, but RNAi and apoptosis also have demonstrated antiviral activity (Clem, 2001; Bronkhorst *et al.*, 2012; Kemp *et al.*, 2013) and the JAK-STAT is active during infection, possibly mediating a tolerance response (West and Silverman, 2018). Baculovirus, nudivirus, and iridovirus infections of *Drosophila* all give rise to virus-derived small interfering RNA (viRNAs), which regulate DNA virus gene expression (Bronkhorst *et al.*, 2012; Jayachandran, Hussain and Asgari, 2012; Kemp *et al.*, 2013; Webster *et al.*, 2015) and mutants for RNAi effectors *Dicer-2* (*Dcr2*) and *Argonaute-2* (*Ago2*) are hypersensitive to Insect Iridescent Virus 6 (IIV6; an iridovirus) infection. This suggests that RNAi may also be an important defence against DNA viruses, and IIV6 correspondingly encodes a suppressor of RNAi (Bronkhorst *et al.*, 2012; Bronkhorst *et al.*, 2014). Virus-encoded suppressors of apoptosis are also widespread in DNA viruses, acting through binding and inhibition of cellular caspases (e.g. p35), or stabilization of cellular inhibitors of apoptosis (e.g. IAP gene family; Bump *et al.*, 1995; Xue and Robert Horvitz, 1995; Byers, Vandergaast and Friesen, 2016). In contrast, the contribution of transcriptional responses, such as the NF-κB pathways, to DNA viruses has not yet been elucidated.

There are two NF-κB pathways in *Drosophila*: Toll and IMD, which primarily function in antibacterial (Toll: gram-positive, IMD: gram-negative) and antifungal (Toll) defense, although both provide protection against some RNA viruses (reviewed in Valanne, Wang and Ramet, 2011; Merkling and van Rij, 2013; Lamiable and Imler, 2014; Myllymaki, Valanne and Ramet, 2014; Palmer, Varghese and van Rij, 2018). Toll and Imd pathways are activated following recognition of a pathogen-associated molecular pattern (PAMP; e.g. bacterial peptidoglycan), leading to the phosphorylation and degradation of the inhibitor of kappa B (IκB; encoded by *cactus* for Toll signalling, and by the *relish* C-terminus in IMD signalling) (reviewed in Valanne, Wang and Ramet, 2011; Myllymaki, Valanne and Ramet, 2014). Under non-signalling conditions, IκB sequesters NF-κB transcription factors in the cytoplasm. These transcription factors are encoded by dorsal (dl) and Dorsal immune-related factor (Dif) in Toll signalling, and C-terminal relish (rel) in IMD signalling, and all translocate to the nucleus to induce gene expression following IκB degradation (reviewed in Valanne, Wang and Ramet, 2011; Myllymaki, Valanne and Ramet, 2014). Although the mechanism by which Toll and IMD

recognise RNA viruses is unclear, both are active and provide immunity against some viral infections in insects, most likely through induction of antiviral effector responses. For example, Toll is broadly antiviral against RNA viruses such as Drosophila C Virus, Nora Virus, and Flock House Virus in *Drosophila* during orally acquired, but not systemic infections, and in *Aedes* mosquitoes against dengue virus (Zambon *et al.*, 2005; Xi, Ramirez and Dimopoulos, 2008; Ramirez and Dimopoulos, 2010; Ferreira *et al.*, 2014). Additionally, Imd is antiviral against a subset of viruses in *Drosophila*, such as Cricket Paralysis Virus, Drosophila C Virus, and Sindbis virus and in *Aedes* cell culture against the alphaviruses Semliki Forest virus and O'nyong'nyong virus (Fragkoudis *et al.*, 2008; Avadhanula *et al.*, 2009; Costa *et al.*, 2009; Waldock, Olson and Christophides, 2012; Sansone *et al.*, 2015).

Although the effect of NF-κB signalling on DNA virus infection in insects has not been directly tested, polydnaviruses, ascoviruses, baculoviruses, and entomopoxviruses have acquired suppressors of NF-κB signalling by horizontal gene transfer, providing indirect evidence for anti-DNA virus NF-κB signalling (Thoetkiattikul, Beck and Strand, 2005; Lamiable, Kellenberger, *et al.*, 2016). First, a 'polydnavirus' encoded by the Braconid wasp *Microplitis demolitor* has also acquired homologs of IκB, some of which inhibit Dif and rel by direct binding (Thoetkiattikul, Beck and Strand, 2005). However, this is a domesticated endogenous viral element that forms a component of the wasp venom, and as these IκB homologues are not found in related nudiviruses, baculoviruses, or hytrosaviruses, it seems likely they were acquired to inhibit anti-parasitoid immune responses in the wasp host rather than wasp antiviral immunity (Bitra, Suderman and Strand, 2012; Herniou *et al.*, 2013). Second, homologs of *diedel* (*die*), a cytokine that inhibits the Imd pathway in *Drosophila*, are similarly found in ascoviruses, baculoviruses, entomopoxviruses, and *Leptopilina spp.* polydnavirus venom, likely through independent horizontal transfer from arthropod hosts (Lamiable, Kellenberger, *et al.*, 2016). Virus-encoded *die* phenocopies fly-encoded *die*, indicating viral die has retained an Imd-suppressive function, and that the Imd pathway likely interacts with these DNA viruses (Lamiable, Kellenberger, *et al.*, 2016). However, it is still unclear whether antiviral Toll signalling is targeted by insect virus-encoded immune suppressors, and whether these hijacked host pathway inhibitors represent a subset of a greater diversity of NF-κB immune inhibitors or reflect evasion of virus-specific immune mechanisms.

The recent isolation of Kallithea virus (KV; Webster *et al.*, 2015; Palmer *et al.*, 2018), a nudivirus that naturally infects *Drosophila melanogaster* at high prevalence in the wild, provides

a tractable system to study host-DNA virus interactions and to identify immune evasion strategies in DNA viruses. Moreover, because some previously-identified virus-encoded immune suppressors have been found to be highly host-specific, the use of a native host-virus pairing such as this is vital to our understanding of viral immune evasion (e.g. Parisien, Lau and Horvath, 2002; Mariani *et al.*, 2003; Goffinet *et al.*, 2009; Elde *et al.*, 2012; Rajsbaum *et al.*, 2012; van Mierlo *et al.*, 2014; Stabell *et al.*, 2018). Here, we use this system to describe the interaction between antiviral immune pathways and a DNA virus in *Drosophila*. Using mutant fly lines, we find the RNAi and Imd pathways mediate antiviral protection against KV *in vivo*, but that abrogation of Toll signalling has no effect on virus replication. Through re-analysis of previous RNA-sequencing data, we observe a broad downregulation of NF-κB responsive antimicrobial peptides following infection, and perform a small-scale screen for KV-encoded immune inhibitors. We identify viral protein gp83 as having a complex interaction with NF-κB signalling, leading to induction of Imd signalling but potent suppression of Toll signalling. This suppression acts directly through, or downstream of, NF-κB transcription factors. Finally, through deletions of conserved protein regions and analysis of the related *Drosophila innubila* nudivirus (DiNV) gp83 ortholog, we show the immunosuppressive activity of gp83 against *D. melanogaster* NF-κB signalling is conserved.

5.3 Materials and Methods

5.3.1 Fly strains, virus growth, and mortality experiments

All fly lines were maintained and crossed on standard cornmeal medium at 25 °C. We measured viral titre and mortality following KV infection in two control lines (w^{1118} and *Oregon R*) and for mutant lines compromised in the following immune signalling pathways: RNAi ($Dcr\text{-}2^{L811fsX}$ (Lee *et al.*, 2004) and $Ago\text{-}2^{414}$ (Okamura *et al.*, 2004)), Toll (*spz,* dl^{1} (Nusslein-Volhard, 1979)*,* Dif^{1} (Rutschmann *et al.*, 2000)*, and* pll^{2}/pll^{21} *trans-heterzygotes* (Anderson and Nüsslein-Volhard, 1984; Hecht and Anderson, 1993)), and Imd (rel^{e20} (Hedengren *et al.*, 1999) *and* Imd^{10191} (Pham *et al.*, 2007)).

For mortality assays, we injected 100 female flies of each genotype with 50 nL of either KV suspension (10^5 ID50) or chloroform-treated KV suspension (which inactivates KV through the destruction of the membrane) and transferred flies in groups of 10 to sucrose agar vials. We recorded the number of surviving flies on alternate days, transferring each group of flies

to fresh food each week. We used the Bayesian generalised linear mixed modelling R package, MCMCglmm (Hadfield, 2010), to analyse per-day mortality as a binomial response variable with days post-inoculation (DPI), DPI^2 (to allow for non-linear changes in mortality), and genotype as fixed effects, and vial as a random effect, as described previously (Palmer *et al.*, 2018). Mortality in RNAi and NF-κB mutants were assayed in separate experiments, and therefore analysed independently. We reported all confidence intervals as 95% from GLMMs as highest posterior density (HPD) intervals.

We measured viral titre in each line after intrabdominal injection of 50 nL of KV suspension. We transferred 50 infected female flies from each line to 10 sucrose/agar vials in groups of 5, and homogenised 5 vials of each genotype in Trizol at 5 and 10 DPI. For RNAi mutants, we also assayed flies 3 DPI. We extracted DNA by phenol-chloroform precipitation and measured viral titre relative to rpl32 by quantitative PCR, using previously described primers (Palmer *et al.*, 2018). We analysed log-transformed viral titre as a Gaussian response variable using MCMCglmm (Hadfield 2010), with genotype, DPI, and genotype-by-DPI interactions as fixed effects. To account for the impact of differing genetic backgrounds between mutant lines, we took a statistical approach using the range seen previously across 120 different natural genetic backgrounds from the *Drosophila* Genetic Reference Panel (Palmer *et al.*, 2018). Specifically, considering w^{1118} and *Oregon R* as controls and mutants of each pathway as the 'experimental' group, we created a null distribution of effect sizes expected only from differences in genetic background, by randomly choosing two DGRP lines to serve as controls and additional DGRP lines reflecting the mutant lines used in each pathway. For each null draw we fitted the same model as described above, recording the absolute value of the effect size, and we repeated the null draw 1000 times to obtain a distribution. If the average effect size associated with mutants in a pathway was greater than the highest 5% of effect sizes, we concluded the observed differences in KV titre were due to mutations in the tested pathway.

5.3.2 Cell culture and virus propagation

We cultured S2 cells at 25 C in Schneiders Drosophila Medium with 10% heat-inactivated fetal bovine serum and 50 U/mL penicillin and 50 ug/mL streptomycin (Life technologies). KV was purified from flies 10 days after initial infection as previously described (Palmer *et al.*, 2018).

5.3.3 Cloning

We selected 9 KV genes identified as highly expressed three DPI (Palmer *et al.*, 2018) to screen for KV-encoded immune suppressors. These were *gp23, gp43, gp83, ACH96233.1-like, ACH96143.1-like, putative protein 1, putative protein 12, putative protein 15, putative serine protease* (corresponding to GenBank accession numbers AKH40365.1, AKH40394.1, AKH40369.1, AKH40392.1, AKH40340.1, AQN78560.1, AKH40392.1, AKH40404.1, and AQN78556.1). Each KV gene was amplified using Qiagen Long Range PCR kit as per the manufacturer's instructions, with primers that introduced restriction sites and the *Drosophila* Kozak sequence (restriction enzymes and primers used in Supplementary Table 1), and cloned into a pAc5.1 vector (Invitrogen) with a V5-His C-terminal tag. The KV gene *gp83* was also cloned into pAc5.1 vector with GFP instead of V5-His to introduce a C-terminal GFP tag. We used PCR splicing to create gp83 deletions ($gp83^{\Delta 1}$: CGLIECSELLRDRLCSKL deletion; $gp83^{\Delta 2}$: WSDRLNLI deletion) by separately amplifying 2 segments of gp83 with primers that span the desired deletion and used these segments as a template for a second PCR reaction (Supplementary Table 1). The resulting amplicons with deletions were cloned as described above. The gp83 gene from DiNV was also cloned as above (Supplementary Table 1).

Additionally, we cloned the Toll pathway components *pll*, *tube, cact, Dif*, and *dl* into the pAc5.1 vector, as described above (Supplementary Table 1). Other Toll and Imd pathway constructs have been described before, including pAc5-$Toll^{LRR}$ (Tauszig *et al.*, 2000), pAc5-dl^{GFP} (Li and Dijkers, 2015), pAc5-PGRP-LC (Kaneko *et al.*, 2006), pAc5-rel^{GFP} (Foley and O'Farrell, 2004), and the firefly luciferase (FLuc) reporter plasmids with promoter sequences from *Drosomycin* (*Drs*), *Diptericin* (*Dpt*), and *Attacin-A* (*Att-A*) (Tauszig *et al.*, 2000) or with 10X STAT binding sites (a gift from Norbert Perrimon, Addgene plasmid # 37393; Baeg, Zhou and Perrimon, 2005).

5.3.4 Transfection and RNAi Knockdown of S2 cells

S2 cells were transfected using Effectene transfection reagent, as per manufacturer's instructions. We synthesized dsRNA against cactus, gp83, FLuc, renilla luciferase (RLuc), and GFP to knockdown these genes in S2 cells. Primers with flanking T7 sequences were used to amplify regions of each gene (Supplementary Table 1) and dsRNA was synthesized from the resulting PCR products with T7 RNA polymerase and purified using GenElute Total RNA mini kit (Qiagen).

5.3.5 Immune suppression assays

We tested whether the 9 cloned KV genes were able to suppress RNAi, JAK-STAT, Toll, or Imd activity. We performed RNAi suppression assays as described previously (van Cleef *et al.*, 2011). Briefly, we seeded $5x10^4$ S2 cells in a 96 well plate and 24 hours later transfected with 33 ng of pMT-FLuc, 33 ng pMT-Rluc, and 33 ng of either pAc5 empty vector or the pAc5 expression construct containing a KV gene. Two DPI, we added 400 ng of either GFP or GL3 dsRNA to each well, and 8 hours later we added $CuSO_4$ to a final concentration of 500 µM to induce expression of the luciferase constructs. We measured RLuc and FLuc luciferase activity using the Dual Luciferase Assay Kit (Promega).

For JAK-STAT immunosuppression assays, we seeded $5x10^4$ S2 cells in a 96 well plate and transfected 24 hours later with 30 ng of 10XSTAT-FLuc, 20 ng pAc5-Rluc, and 50 ng of either pAc5 empty vector or the pAc5 expression construct containing a KV gene. We measured luciferase readings 48 hours following transfection.

For NF-κB immunosuppression assays, we transfected either a construct containing Imd receptor pMT-PGRP-LCx or a constitutively active Toll construct lacking the extracellular leucine-rich repeat domain, pAc5-Toll^{LRR} (Tauszig *et al.*, 2000)alongside each KV gene, and a NF-κB-responsive FLuc reporter containing the either the *Dpt* (Imd) or *Drs* (Toll) promoter sequence (Tauszig *et al.*, 2000). We seeded $5x10^4$ S2 cells in 96 well plates and 24 hours later transfected 50 ng of either empty pAc5.1 vector or a pAc5.1 KV gene expression construct, 20 ng of either pAc5.1 or pAc5.1-Toll^{LRR}, 10 ng of Drs-FLuc, and 10 ng pAc5-Rluc. We performed analogous experiments using pAc5.1-dl, pAc5.1-Dif, and pAc5.1-pll instead of pAc5.1-Toll^{LRR}, or by transfecting 5 ng of *cact* dsRNA. In the latter, 70 ng of KV gene expression construct was transfected instead of 50 ng. We assayed RLuc and FLuc activity 48 hours after transfection.

We also performed immunosuppression assays using KV-infected cells. We seeded $5x10^4$ cells in 96 well plates and immediately added 5 uL of KV suspension (10^3 ID50), or chloroform-treated KV, transfected cells the next day. For RNAi suppression assays with KV, we transfected 50 ng pMt-RLuc, 50 ng pMT-FLuc, and 5 ng of either GFP or GL3 dsRNA 2 days after infection and added $CuSO_4$ 8 hours later. To measure JAK-STAT activity following KV infection, we transfected 70 ng of 10XSTAT-FLuc and 30 ng pAc5-Rluc. For Toll suppression assays, we transfected 70 ng of either pAc5 or pAc5-Toll^{LRR}, 20 ng of Drs-FLuc, and 10 ng pAc-RLuc. Finally, to measure Imd activity following KV infection, we transfected 70 ng of either pMT or

pMT-PGRP-LCx, 20 ng of Dpt-FLuc, and 10 ng pAc-RLuc, and added CuSO4 immediately following transfection. We measured luciferase readings 4 DPI.

We used the R package MCMglmm to determine significance in immune suppression assays, with the RLuc-normalised FLuc values as a Gaussian response variable. In the original screen for immune suppressors, we treated any experimental induction of an immune pathway as a fixed effect (e.g. addition of dsRNA against FLuc in the RNAi suppression assay, PGRP-LC overexpression in the Imd suppression assay, and $Toll^{LRR}$ in the Toll suppression assay), each KV gene as a random effect, and the interaction between KV gene and the induced experimental change to signalling output as a random effect. In subsequent NF-κB suppression experiments, where the only tested KV gene was gp83, we treated gp83 and the interaction between gp83 and overexpression of NF-κB receptors as fixed effects. Likewise, when immune suppression experiments were carried out with KV-infected cells instead of over-expressing KV genes, we treated KV infection status, the induction of an immune pathway, and the interaction between these as fixed effects. We have made all code used to fit the models, and associated data available on Figshare (doi: 10.6084/m9.figshare.c.4151009).

5.3.6 Immunoprecipitation and western blotting

We tested whether gp83 directly interacted with dl by seeding $2x10^6$ cells in 6 well plates and transfecting 150 ng of either pAc5.1 empty vector, V5-tagged gp83, or V5-tagged cact and 150 ng of either GFP of GFP-tagged dl. Two days post-transfection, we resuspended two wells per treatment in lysis buffer (0.1% NP-40, 30 mM Hepes-KOH, 150 mM NaCl, 2mM MgOAc) supplemented with protease inhibitor and 5 mM DTT, and passed the cell lysate through a 25-gauge needle 30 times. After 10 minutes incubation on ice, we pelleted cell debris by centrifuging at 16000xg for 30 minutes and supernatant was either stored as an input control or collected and incubated for 5 hours at 4 C with magnetic control beads. We removed binding control beads and incubated the resulting supernatant with GFP-trap (Chromotek) magnetic beads overnight at 4 degrees. We washed beads 3 times in lysis buffer and 3 times in 25 mM Tris-HCl, 150 mM NaCl solution, and eluted protein complexes by boiling 10 minutes at 95 C in Laemmli buffer.

We prepared whole cellular protein extracts by boiling S2 cells for 10 min at 95 C in Laemmli buffer. We separated whole cellular extracts or immunoprecipitated proteins on an SDS-PAGE gel and transferred to a nitrocellulose membrane. We blocked non-specific binding with blocking solution (PBT with 5% dry milk), probed for proteins of interest with primary

antibody diluted in blocking solution overnight at 4 C, and visualized proteins with an hour incubation of secondary antibody in blocking solution. We washed membranes 3 times in PBT before and after each step. We used mouse anti-dl (1:100, Developmental Studies Hybridoma Bank), mouse anti-V5 (1:1000 dilution, Invitrogen), rat anti-tub-α (1:1000 dilution, SanBio), and rabbit anti-GFP (abcam ab6556) as primary antibodies, and goat anti-mouse IR-Dye 680 (1:15000, LI-COR), goat anti-rat IR-Dye 800 (1:15000, LI-COR), goat anti-rabbit IR-Dye 800 (1:15000, LI-COR). We used an Odyssey Infrared Imager (LI-COR) to image blots.

5.3.7 Mass spectrometry

We co-transfected 10^6 S2 cells with pCoBLAST and $GP83^{GFP}$ plasmid at a 1:19 ratio (125ng and 2.38µg, respectively). We replaced medium 3 hours post-transfection, and again at 48 hours post-transfection with medium supplemented with blasticidin (20µg/mL). Another 48 hours later, we refreshed the cells with medium containing 4µg/mL blasticidin, which was thereafter replaced every 3-4 days with medium containing 4µg/mL blasticidin, resulting in a polyclonal cell line.

For mass spectrometry, we lysed either wild-type S2 cells or S2 cells stably expressing $GP83^{GFP}$ in lysis buffer containing 50mM Tris-HCl (pH 7.8), 150mM NaCl, 1% NP-40, 0.5mM DTT, 10% glycerol and protease inhibitor cocktail (Roche). Approximately 4 mg protein lysate was subjected to GFP-affinity purification using 7.5uL GFP-TRAP beads (Chromotek) for ~1.5 hours at 4ºC. Beads were washed twice in lysis buffer, twice in PBS containing 1% NP-40, and three times in PBS, followed by on-bead trypsin digestion as described previously (Smits *et al.*, 2013). Afterwards, tryptic peptides were acidified and desalted using Stagetips, eluted, and brought onto an EASY-nLC 1000 Liquid Chromatograph (Thermo Scientific). Mass spectra were recorded on a QExactive mass spectrometer (Thermo Scientific) and MS and MS2 data were recorded using TOP10 data-dependent acquisition. We used Maxquant (v1.5.1.0) to analyse raw data, using recommended settings (Cox and Mann, 2008). LFQ, IBAQ, and match between runs were enabled. We mapped the peptides to *Drosophila melanogaster* proteins (UniProt June 2017) and used Perseus (v1.3.0.4) to filter contaminants and reverse hits. We imputed missing values, assuming a normal distribution and performed a t-test on log-transformed LFQ-values to calculate significantly enriched proteins.

5.3.8 Immunofluorescence microscopy

We seeded $5x10^5$ S2 cells in 12-well plates with glass coverslips in each well. We transfected cells with 100 ng of pAc5.1 or pAc5.1-gp83^{V5} and 100 ng of dI^{GFP}. Two days after transfection, we fixed cells with 4% paraformaldehyde, washed twice in PBS, once with PBT, and blocked with PBT with 10% goat serum. We stained with mouse anti-V5 (1:400, Invitrogen) for one hour, followed by fluorophore-containing goat anti-mouse secondary antibody (1:400, LI-COR) with 10 ug/mL DAPI for one hour. We washed cells twice in PBT and twice in PBS, then mounted them on slides with Fluoromount-G (eBiosciences). We imaged slides with an Olympus FluoView FV1000.

5.4 Results and Discussion

5.4.1 RNAi and Imd pathways are antiviral against KV in vivo

The RNAi pathway provides antiviral activity against the DNA virus IIV6 and Ago2 is upregulated in response to KV, while the contribution of Imd and Toll pathways to anti-DNA virus immunity have not been described (Bronkhorst *et al.*, 2012; Alfred W Bronkhorst *et al.*, 2014). We used fly lines mutant for RNAi, Imd, and Toll pathway components to assess whether these pathways fulfil an antiviral function during KV infection. First, we infected mutants for RNAi genes *Dcr2* and *Ago2* with KV, and measured viral titre and mortality following infection. Following KV infection, both *Dcr2* and *Ago2* mutants exhibited significantly greater KV titres at 3 DPI, with KV titre 78-fold greater in *Dcr2* mutants (95% HPD intervals: 18-281 fold; MCMCp < 0.001) and 55-fold greater in *Ago2* mutants (13-237 fold, MCMCp < 0.001; Figure 1). However, the increased KV replication in RNAi mutants was not sustained at later infection timepoints. At 5 DPI, *Dcr2* mutants did not have significantly different KV titre from the controls (MCMCp = 0.22), and KV had a slightly diminished advantage in *Ago2* mutants (12-fold increase; 2.5-43, MCMCp < 0.001; Figure 1). By 10 DPI, there was no significant difference between viral titre in control flies and either *Dcr2* mutants (MCMCp = 0.43) or *Ago2* mutants (MCMCp = 0.7). Therefore, either the antiviral effect of RNAi is short-lived (i.e. a viral suppressor of RNAi is eventually expressed *in vivo*), other immune pathways take over as the dominant antiviral force, KV negatively regulates its own replication or saturates a resource. Nevertheless, despite the similar titres during late infection, there was still a significant increase in KV-induced mortality in *Dcr2* and *Ago2* mutants, where 70% of control flies were alive 19 DPI, compared to 25% in *Dcr2* mutants (MCMCp < 0.001) and 38% in *Ago2* mutants

(Figure 1), possibly due to early host damage or increased expression of virulence factors throughout infection (e.g. Jayachandran, Hussain and Asgari, 2012). These results extend the antiviral role of the RNAi pathway to KV infection.

We next infected Imd and Toll pathway mutants with KV, and assessed KV titre by qPCR at 5 and 10 DPI. We found Imd pathway mutants had significantly greater viral titre as compared to two control lines, with *imd* mutants having 6-fold greater KV titre at 5 and 10 DPI (2.7-13.7-fold, MCMCp < 0.001), and *rel* having 8-fold greater viral titre at 5 and 10 DPI (3.1-15.9-fold, MCMCp < 0.001; Figure 1). Because the Imd effect spans 5 and 10 DPI, and we have previously measured KV titre in 125 inbred lines of the *Drosophila* Genetic Reference Panel at 8 DPI (Mackay *et al.*, 2012; Palmer *et al.*, 2018), we attempted to account for genetic background by comparing the average effect of Imd mutants to the distribution of effects consistent with natural variation in the genetic background. This analysis indicated that the increased titre observed in Imd mutants is unlikely to be due to genetic background ($p = 0.01$), if the variation in the DGRP is representative of the variation expected between lab-maintained fly lines. We also infected flies mutant for the Toll pathway components *spz*, *pll*, *Dif*, and *dl.* Viral titre was unchanged in most Toll pathway mutants compared to controls, except in *Dif* mutants (MCMCp = 0.02; Figure 1), where some vials remained uninfected. However, we note that the *Dif* mutant individuals were much larger than those of other lines, and therefore the decreased titre could be due to differences in body size. Further, the pathway-level effect of Toll mutants was within the expected distribution of effects caused by differences in genetic background, even when *Dif* mutants were excluded ($p = 0.28$). We concluded that the Imd pathway is antiviral against KV, but that abrogation of Toll function has no effect on KV growth. This could indicate that Toll is not antiviral against this DNA virus, or that it is efficiently suppressed by virus infection. The latter is consistent with our observation that antimicrobial peptides are generally downregulated during KV infection (Figure 1), and we therefore explored the capability of KV to suppress innate immune pathways in cell culture.

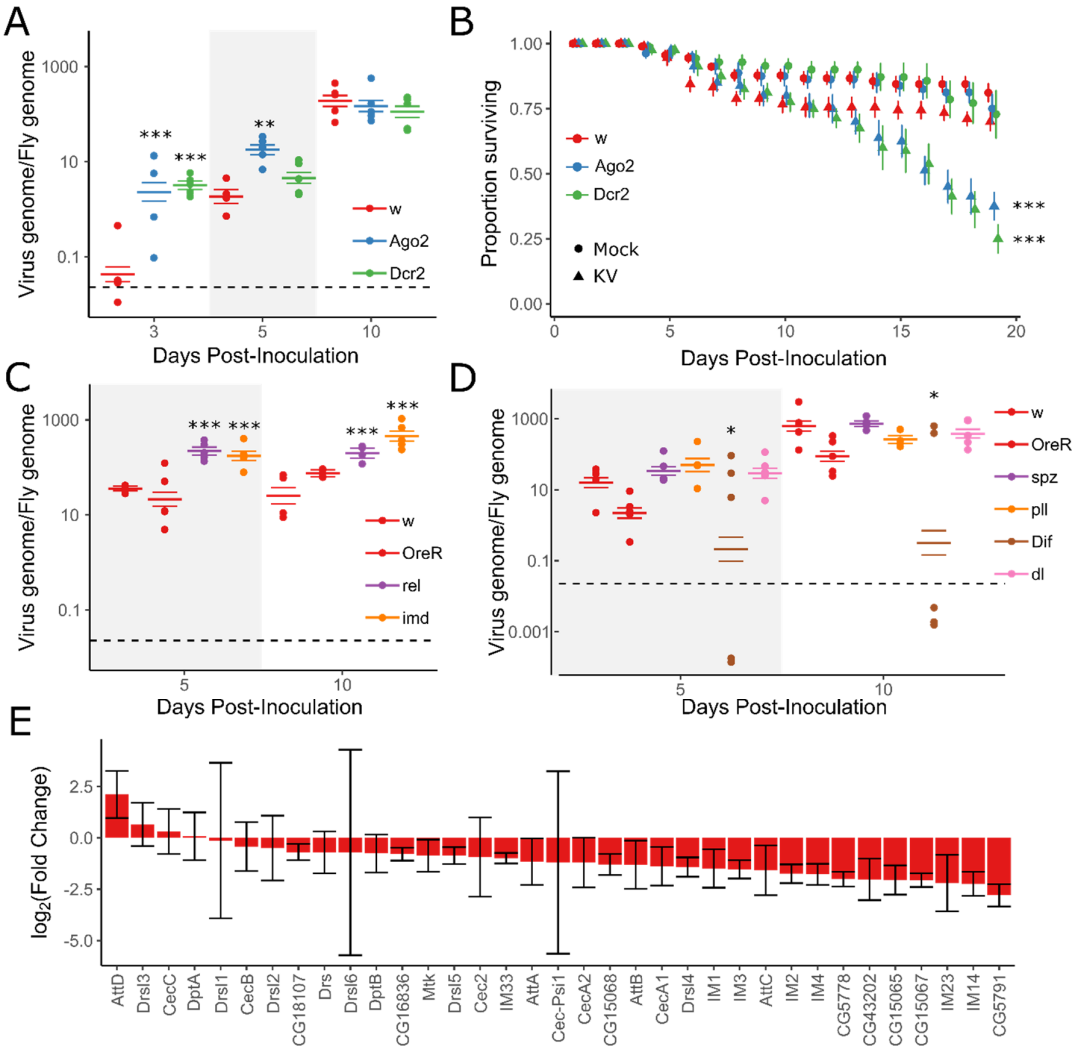

Figure 1: RNAi and Imd pathways provide antiviral defense against Kallithea virus

We infected mutants for RNAi (A,B) and NF-κB (C,D) pathways and assayed viral titre and mortality following infection. Horizontal dotted lines (A,C,D) represent the amount of virus injected at the zero timepoint. (A) RNAi mutants have increased early viral titre relative to control w^- flies at 3 DPI, but no difference in titre by 10 DPI. (B) Despite similar titres during late infection, RNAi mutants exhibit increased mortality following KV infection, relative to chloroform-treated KV controls. (C) Imd and rel mutant flies have increased viral titre relative to two wild-type lines (w^- and OreR) at 5 and 10 DPI, however (D) Toll pathway mutants spz, pll, Dif, or dl show no consistent difference in KV titre relative to control lines at 5 or 10 DPI. (E) Toll and Imd-responsive antimicrobial peptides are generally downregulated following KV infection (Data from Palmer et al, 2018; ERP023609), consistent with the presence of a virus-encoded immune inhibitor. Error bars show standard errors of the mean.

5.4.2 KV growth in cell culture

KV growth in *D. melanogaster* cell culture has not previously been described. We found variation in the ability of KV to infect five commonly-used cell lines (Figure S1), such that KV grew well in S2, S2R+, and DL2 cells but poorly in Kc167 and Dm-BG3-c2 cells. In S2 cells, which we

used for further analyses, KV was released into the medium at substantial levels starting from 3 DPI (Figure S1). Therefore, in all subsequent experiments, we assayed cells at 4 DPI, assuming that a high proportion of cells would be infected at this timepoint. We did not observe any overt cytopathic effects of KV-infected cells within 14 days of infection. However, when KV-infected cells were passaged 7 DPI, we observed larger (MCMCp < 0.001) and fewer (MCMCp < 0.001) cells, likely due to a decrease in cell proliferation.

5.4.3 KV inhibits JAK-STAT and Toll, and induces Imd signalling in cell culture

We used previously established luciferase reporter-based assays to describe the effect of KV infection on RNAi, JAK-STAT, Toll, and Imd pathways. To determine if KV suppresses RNAi, we measured the RNAi silencing efficiency of cells treated with KV or chloroform-inactivated KV (hereafter referred to as mock-treated) by co-transfecting FLuc and with either GFP dsRNA or FLuc dsRNA. In both mock and KV-treated cells, FLuc dsRNA caused a 95% reduction in FLuc activity compared with GFP dsRNA treated cells, indicating KV infection does not inhibit RNAi in cell culture (MCMCp = 0.9; Figure 2). Most viruses studied in *Drosophila* encode a VSR (e.g. Li, Li and Ding, 2002; Van Rij *et al.*, 2006; Nayak *et al.*, 2010; van Mierlo *et al.*, 2012, 2014; Bronkhorst *et al.*, 2014), and therefore the absence of KV-induced RNAi suppression is somewhat surprising. It is possible that KV-RNAi interactions are different in the cell type KV naturally infects, and the inability to observe an effect is a limitation of the cell culture model. Alternatively, if KV transmission does not occur until later stages of infection, there may be limited selective pressure to evade RNAi, as RNAi mutants and control flies have similar titres during late infection.

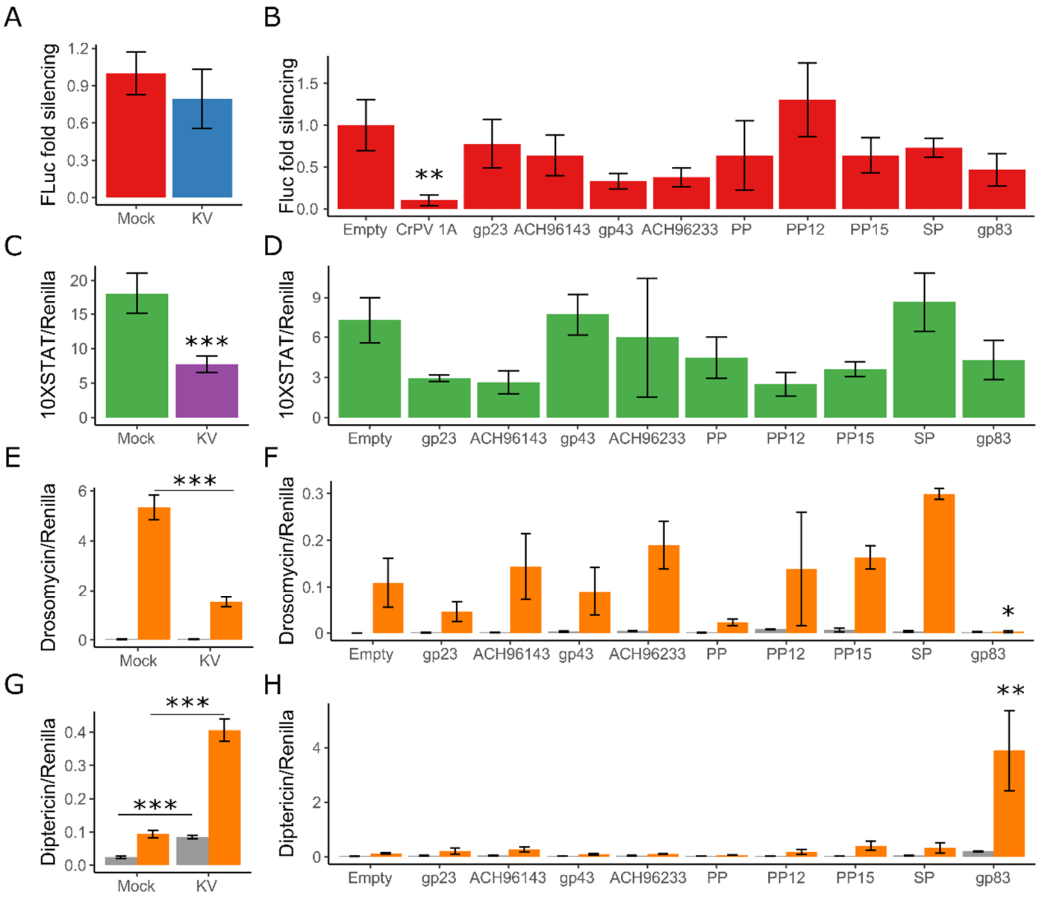

Figure 2: Kallithea virus interactions with innate immune pathways

We assessed the ability of KV to inhibit RNAi, JAK-STAT, Toll, and Imd pathways (A,C,E,G), and whether 9 highly expressed KV genes interacted with these pathways (B,D,F,H). Chloroform-treated KV are denoted "Mock" infections (A,C,E,G) and transfection with the plasmid used to clone genes denoted "Empty" (B,D,F,H). (A) KV was unable to effectively inhibit RNAi silencing, as incubation with FLuc dsRNA was able to efficiently silence FLuc during KV infection. (B) Additionally, no tested KV gene inhibited RNAi, although overexpression of protein 1A from Cricket Paralysis virus was able to almost completely inhibit silencing (data combined from 2 independent experiments). (C) KV infection reduced JAK-STAT signalling in S2 cells, however (D) overexpression of any of the 10 highly expressed KV genes was not significantly associated with STAT suppression. (E) Toll is not endogenously active in S2 cells (grey bars) but overexpression of Toll^{LRR} resulted in a dramatic increase in Drs luciferase (orange bars), which KV partially inhibited. (F) Overexpression of gp83 was able to completely inhibit Toll-induced Drs expression. (G) Likewise, Imd is not active in S2 cells, and overexpression of PGRP-LC led to increased Dpt luciferase (compare grey and orange bars). KV infection significantly induced Dpt luciferase with or without PGRP-LC. (H) Additionally, gp83 overexpression potently induced Imd signalling when coupled with PGRP-LC overexpression. PP=Putative Protein; SP=Serine Protease. Error bars show standard errors of the mean, calculated from 5 biological replicates for (A,C,E,G) and 3 biological replicates for (B,D,F,H).

The JAK-STAT pathway has an antiviral role during Drosophila C Virus infection (Dostert *et al.*, 2005) and mediates tolerance to the DNA virus IIV6, evidenced by upregulation of vir-1 and the Turandot (Tot) genes (West and Silverman, 2018). However, previous transcriptional profiling did not identify differential expression of these genes following KV infection (Palmer *et*

al., 2018). We assessed JAK-STAT activity in mock and KV-treated cells with a FLuc reporter containing ten STAT binding sites (Baeg, Zhou and Perrimon, 2005). This reporter is endogenously active in S2 cells, but KV infection led to a 58% reduction in STAT-mediated FLuc activity (37-74%, MCMCp < 0.001; Figure 2), indicating JAK-STAT is down-regulated or inhibited following KV infection, and suggesting the differences between JAK-STAT involvement in KV and IIV6 infection could be under viral control. However, in addition to mediating a transcriptional immune response, the JAK-STAT pathway is involved in cell proliferation (Arbouzova and Zeidler, 2006), which also decreases following KV infection in cell culture (Figure S1), making cause and effect difficult to distinguish. Nevertheless, vertebrate-infecting viruses set a precedent for virus-mediated JAK-STAT inhibition by diverse mechanisms, including interference with ligand-receptor binding, and downregulation (via miRNAs), degradation, or inhibition of phosphorylation of JAK and STAT (reviewed in Nan, Wu and Zhang, 2017).

We also assayed the effect of KV on Toll and Imd signalling. However, these pathways are not active in S2 cells. To measure KV suppression of these pathways, we therefore co-transfected $Toll^{LRR}$ or PGRP-LC with *Drs* or *Dpt* luciferase reporters to artificially induce signalling of Toll and Imd pathways, respectively. Transfection of $Toll^{LRR}$ increased Drs-Fluc by 243-fold (MCMCp < 0.001) and KV infection reduced the maximum level of $Toll^{LRR}$-mediated Drs activity by 81% (38-93%, MCMCp < 0.001; Figure 2), indicating KV can inhibit Toll signalling. Overexpression of PGRP-LC led to a 4-fold increase in Dpt-FLuc (3-5 fold, MCMCp < 0.001). Similarly, KV infection led to a 3.6-fold increase (2.6-4.8 fold, MCMCp < 0.001) in Dpt-FLuc, which additively increased Dpt-FLuc when PGRP-LC overexpressing cells were infected with KV (17-fold increase compared to Imd-inactive, mock-treated cells; 12-23 fold, Figure 2). This suggests that KV infection in S2 cell culture suppresses Toll signalling and induces Imd signalling.

5.4.4 KV-encoded gp83 interacts with NF-κB signalling during infection

The immunosuppressive function of nudivirus genes has not previously been explored in detail. To identify KV-encoded immune inhibitors, we cloned 9 KV genes that are highly expressed 3 days post infection in adult flies (Palmer *et al.*, 2018) and performed immune suppression assays for RNAi, JAK-STAT, Toll, and Imd pathways. We were unable to identify KV-encoded suppressors of RNAi or JAK-STAT among these 9 genes, although we confirmed that Cricket Paralysis Virus protein 1A potently suppressed RNAi (Nayak *et al.*, 2010; MCMCp = 0.006; Figure 2). However, we found that gp83 – a KV gene encoding no recognisable protein

domains, significantly reduced Toll^{LRR}-induced Drs-FLuc expression (Figure 2). In this experiment, Toll^{LRR} overexpression induced Drs-FLuc by 24-fold (8-66 fold), but only by 1.9-fold (0.3-8 fold; MCMCp = 0.02) when gp83 was co-expressed. We confirmed this was not an artefact of the luciferase reporter-based assay by repeating the experiment with qPCR of endogenous *Drs* as a readout, where gp83 overexpression potently reduced Toll^{LRR}-induced *Drs* expression (MCMCp < 0.001; Figure S2). We further found that overexpression of gp83 caused a 5-fold (1.5-18 fold) *increase* in Imd-mediated Dpt-FLuc expression, with or without PGRP-LC overexpression (MCMCp = 0.008; Figure 2).

We next aimed to confirm the interactions between the transfected KV gene gp83 and NF-κB pathways are representative of the function of gp83 during KV infection. Therefore, we silenced gp83 during KV infection using dsRNA, and measured associated changes in Toll, Imd, and JAK-STAT signalling. Co-transfection of gp83 with independent dsRNAs targeting gp83 completely reversed inhibition of Drs-FLuc compared with transfection of GFP dsRNA, indicating these dsRNAs effectively silence gp83 (MCMCp < 0.001 for both dsRNAs; Figure S3). KV infection had no effect on Drs-FLuc (MCMCp = 0.26), but inhibited Toll^{LRR}-induced signalling (MCMCp < 0.001), as reported above (Figure 2). However, knockdown of gp83 during KV infection led to increased Drs-FLuc in Toll-inactive cells (MCMCp = 0.004) and cells expressing Toll^{LRR} (MCMCp < 0.001; Figure 3). Likewise, knockdown of gp83 in KV-infected cells caused a modest decrease in Dpt-FLuc in Imd-inactive cells (MCMCp = 0.03), and this effect was much stronger in PGRP-LC overexpressing cells (MCMCp = 0.006; Figure 3). Consistent with a specific interaction with NF-κB signalling, gp83 knockdown had no effect on the ability of KV to suppress JAK-STAT signalling (MCMCp = 0.63; Figure 3).

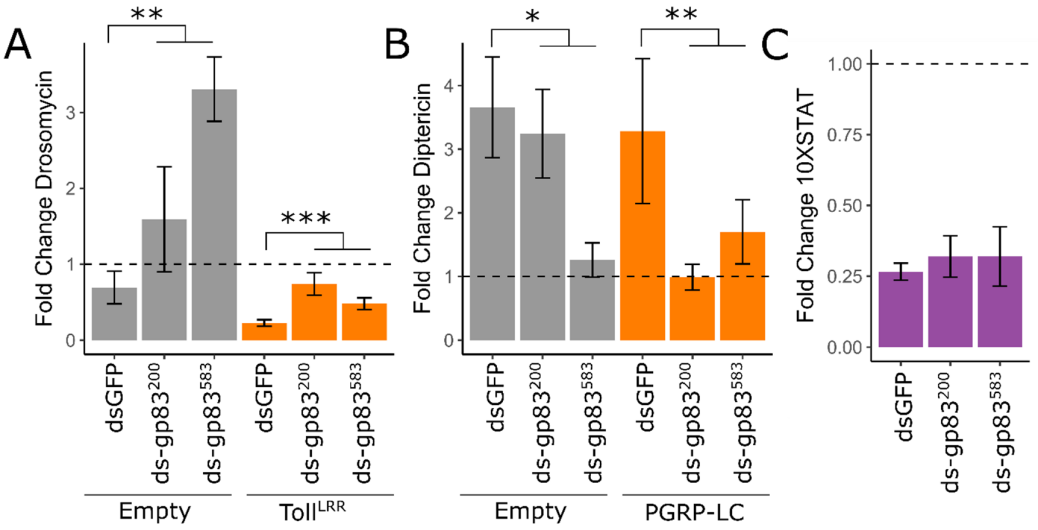

Figure 3: KV induction and suppression of NF-KB pathways is mediated by gp83

We assayed the ability of KV to inhibit Toll (A), induce Imd (B), and inhibit JAK-STAT (C) during gp83 knock down, using two independent dsRNAs against gp83 (labelled gp83^{200} and gp83^{583}; see Figure S4 for confirmation of gp83 KD). Each assay was performed 4 DPI. For each, fold-change in signalling following KV infection is plotted, where 1 (horizontal dotted line) represents no induction or suppression of the pathway. (A) Knock-down of gp83 caused a significant increase in Drs luciferase expression following KV infection (grey bars) and significantly reduced the extent of KV suppression of Toll signalling during Toll^{LRR} overexpression (orange bars). (B) Knock down of gp83 causes significantly decreased Dpt luciferase expression following KV infection, especially during increased Imd activation (PGRP-LC overexpression; orange bars). (C) Knockdown of gp83 had no effect on the ability of KV to suppress JAK-STAT (Figure 2), indicating gp83 interacts specifically with NF-κB signalling during KV infection. Error bars show standard error of the mean (n = 5).

The immunosuppressive function of gp83 on Toll signalling is consistent with the observed downregulation of AMPs following KV infection *in vivo* and is consistent with the hypothesis that Toll is antiviral and suppressed during infection. However, the induction of antiviral Imd signalling by a single viral protein is peculiar, and it is unclear why KV has not evolved to avoid Imd activation, or suppress it as is done in other insect-infecting DNA viruses (Lamiable, Kellenberger, *et al.*, 2016). Assuming Imd activation is detrimental to virus transmission, this could indicate a trade-off between suppression of Toll and activation of Imd or that gp83 is directly recognised by the fly innate immune system. We conclude that KV-encoded gp83 at least partially mediates complex interactions with NF-κB signalling, including suppression of Toll signalling and induction of Imd signalling.

5.4.5 Immune suppression by gp83 occurs downstream of Toll transcription factors

Previously described polydnavirus-encoded Toll pathway inhibitors imitate IκB, blocking the nuclear entry of NF-κB transcription factors (Thoetkiattikul, Beck and Strand, 2005). Although

the detailed mechanism of interaction between gp83 and Toll signalling is unknown, suppression of Toll^{LRR}-induced signalling indicates gp83 also interferes with intracellular Toll signalling. We therefore performed genetic interaction experiments to narrow down the point in the Toll signalling pathway at which gp83 acts. Drs-FLuc was greatly increased by overexpressing *pll* (240-fold induction of Drs-FLuc; 131-414 HPD), silencing *cact* (75-fold induction of Drs-FLuc; 33-161 HPD), and overexpressing *Dif* (563-fold induction of Drs-FLuc; 317-1002 HPD) or *dl* (459-fold induction of Drs-FLuc; 257-778 HPD). Co-overexpression of gp83 potently reduced Drs-FLuc in each of these scenarios (MCMCp < 0.001 for each) – pll/gp83 co-overexpression led to a 0.55-fold change in Drs-FLuc (0.31-0.99 fold), cact^{dsRNA}/gp83 led to a 1.73-fold change in Drs-FLuc (0.75-3.5 fold), Dif/gp83 led to a 0.86-fold change in Drs-FLuc (0.5-1.5 fold), and dl/gp83 led to a 1.5-fold change in Drs-FLuc (0.9-2.5 fold) relative to baseline Drs-FLuc expression (Figure 4). Additionally, V5 staining of gp83^{V5} revealed that gp83 is a nuclear protein (Figure 4). Together, these results indicate that gp83 either inhibits NF-κB transcription factors, or acts downstream of them.

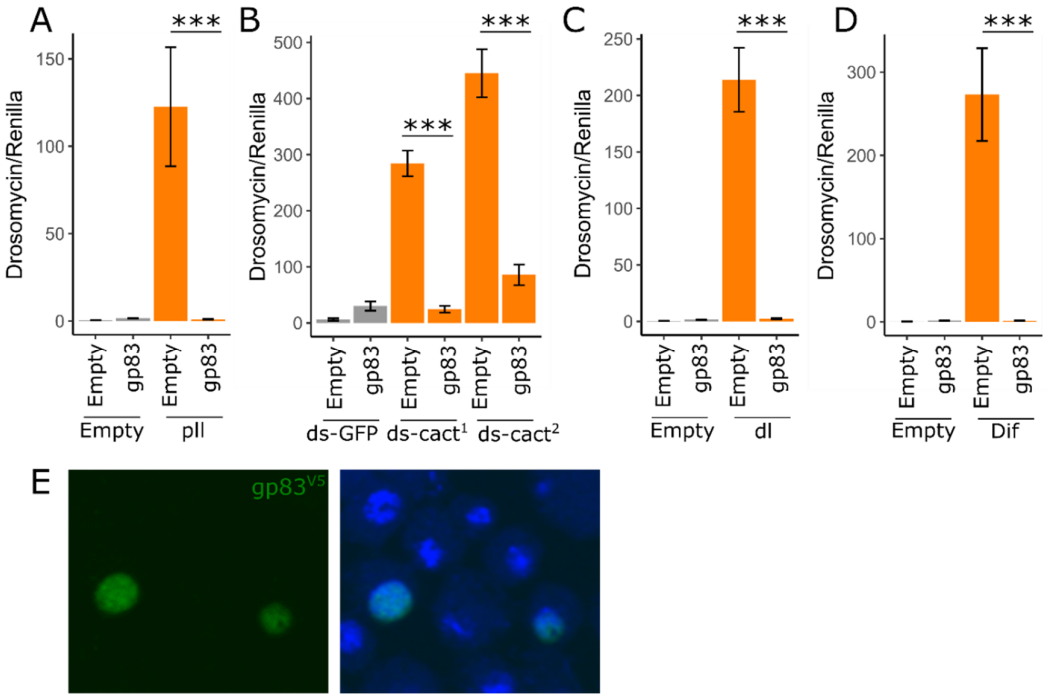

Figure 4: gp83 interacts with Toll signalling downstream of Dif and dorsal

Overexpression of pll (A), knockdown of cactus with two independent dsRNAs (B), and overexpression of NF-κB transcription factors dl and Dif led to increased Drs luciferase expression in S2 cells. In each case, gp83 was able to inhibit signalling, indicating that gp83 inhibited Toll signalling at the level of, or downstream of the NF-κB transcription factors. (E) Consistent with a downstream function, gp83 is a nuclear protein. Error bars show standard error of the mean (n = 5).

Virus-encoded inhibitors of NF-κB in mammals have been reported to operate by promoting degradation of NF-κB transcription factors, blocking NF-κB access to the nucleus, or interfering with transcriptional co-activators to evade the interferon response (reviewed in Zhao *et al.*, 2015). Using co-immunoprecipitation and subsequent western blotting or mass spectrometry, we tested whether gp83 may directly bind and either stabilise or degrade dl. Following immunoprecipitation of GFP-tagged dl, we were able to detect cact as an interacting positive control, but not gp83 (Figure S4). Additionally, we created an S2 cell line stably expressing GFP-tagged gp83, immunoprecipitated gp83GFP, and performed mass spectrometry on interacting partners. We identified 19 Drosophila melanogaster proteins, including 4 nuclear proteins (Nipped-B, Brf, Mlf, Ulp1), that were enriched in the gp83 immunoprecipitate (log2 fold enrichment > 2.5; FDR < 0.1; Figure S4). While we did not identify known downstream NF-κB pathway factors, the extracellular Toll ligand spz was among those enriched, despite the nuclear localization of gp83. However, peptide coverage of spz was poor and dsRNA knockdown of spz did not rescue the immunosuppressive effect of gp83, indicating this interaction may not occur in live cells, or that it is not required for gp83 to inhibit Toll signalling. Further, knockdown of a subset of the enriched genes, including 3 of the 4 identified nuclear proteins, was unable to rescue the gp83 immunosuppressive effect (Figure S4), suggesting that gp83 may not form stable complexes with host proteins to interfere with NF-κB signalling (Figure S4).

Although we did not detect a direct association between dl and gp83, we observed a reduction in dl protein levels upon gp83 overexpression that is not dependent on Toll signalling (Figure S5). We quantified this effect by transfecting either GFP or GFP-tagged dl, with or without gp83, and measuring fluorescence by confocal microscopy. We found that while gp83 caused a 53% reduction in GFP levels, possibly due to a dl binding site in the actin 5C promoter of this construct (Zehavi *et al.*, 2014; 42-62%, MCMCp < 0.001), it caused a significantly greater reduction in dl^{GFP} (73% reduction; 66-78%, MCMCp < 0.001). However, KV infection did not decrease dl protein levels, indicating this is not likely the primary mechanism by which KV inhibits Toll signalling (Figure S5). Instead, we hypothesize that gp83 interferes with the access of dl to either the nucleus or NF-κB binding sites, which indirectly affects dl localization and results in increased turnover. We prefer the latter explanation, that gp83 directly interferes with the Toll pathway transcriptional response, because overexpression of gp83 simultaneously induced the *Dpt* reporter (Figure 2) and reduced dl-responsive promot-

ers (Drs-FLuc and Act5C-GFP; Figure 2, Figure S5). These observations implicate gp83 in regulating transcription at diverse loci responsive to both dl and rel, and suggest an interaction between gp83 and NF-κB-responsive genes.

5.4.6 Immunosuppressive function of gp83 depends on conserved residues and is conserved in other nudiviruses

Conflict between the host immune system and virus-encoded immune inhibitors may be expected to result in an evolutionary arms race, leading to recurrent positive selection and eventual host specialization (e.g. Obbard *et al.*, 2009; Sawyer and Elde, 2012; Brockhurst *et al.*, 2014). Consistent with this, some immune inhibitors are only effective against their native host species, thereby defining the viral host range (e.g. Parisien, Lau and Horvath, 2002; Mariani *et al.*, 2003; Goffinet *et al.*, 2009; Rajsbaum *et al.*, 2012; van Mierlo *et al.*, 2014; Stabell *et al.*, 2018). We tested whether the immunosuppressive function of gp83 is conserved, and whether gp83 acts in a species-specific manner. The gp83 locus is absent from nudiviruses distantly related to KV, such as Heliothis zea nudivirus 1 (HzNV1), HzNV2 and Peneaus monodon nudivirus (PmNV), but is found in more closely related viruses such as Gryllus bimaculatus nudivirus (GrBNV), Nilaparvata lugens endogenous nudivirus (NlENV), Oryctes rhinoceros nudivirus (OrNV), Drosophila innubila nudivirus (DiNV), Tomelloso virus (TV), and Esparto virus (EV). Although gp83 lacks recognisable protein domains, there are regions conserved across these nudiviruses, suggesting functional conservation (Figure 5). We made two gp83 deletion constructs ($gp83^{\Delta1}$ and $gp83^{\Delta2}$) that remove conserved regions of 18 and 8 amino acids without substantially altering protein stability, and transfected these alongside $Toll^{LRR}$ with the Drs-FLuc reporter. Neither $gp83^{\Delta1}$ (MCMCp = 0.67) or $gp83^{\Delta2}$ (MCMCp = 0.79) were able to inhibit Toll signalling, indicating these conserved residues are important for the immunosuppressive function of gp83. Additionally, we cloned gp83 from DiNV and performed Toll immunosuppression assays. The gp83 homolog from DiNV was able to completely inhibit *D. melanogaster* Toll signalling (MCMCp < 0.001), despite only 57% amino acid identity with KV gp83, demonstrating the immunosuppressive function of gp83 is conserved in other nudiviruses but is not highly host-specific. This suggests that the Toll-gp83 interaction may not be hotspot of antagonistic 'arms race' coevolution and has not led to specialization of DiNV gp83 to the *D. innubila* immune system at the expense of its ability to function in *D. melanogaster*. This is also consistent with previous observations that Toll signal transduction genes do not have apparent signatures of positive selection in *Drosophila*

(Sackton *et al.*, 2007; Obbard *et al.*, 2009). This could be because gp83 has very few direct interactions with host proteins (Figure S4), and may instead interact directly with transcription factor binding sites which are under high constraint, and therefore unable to evolve resistance to the immunosuppressive effect of gp83 (Nitta *et al.*, 2015).

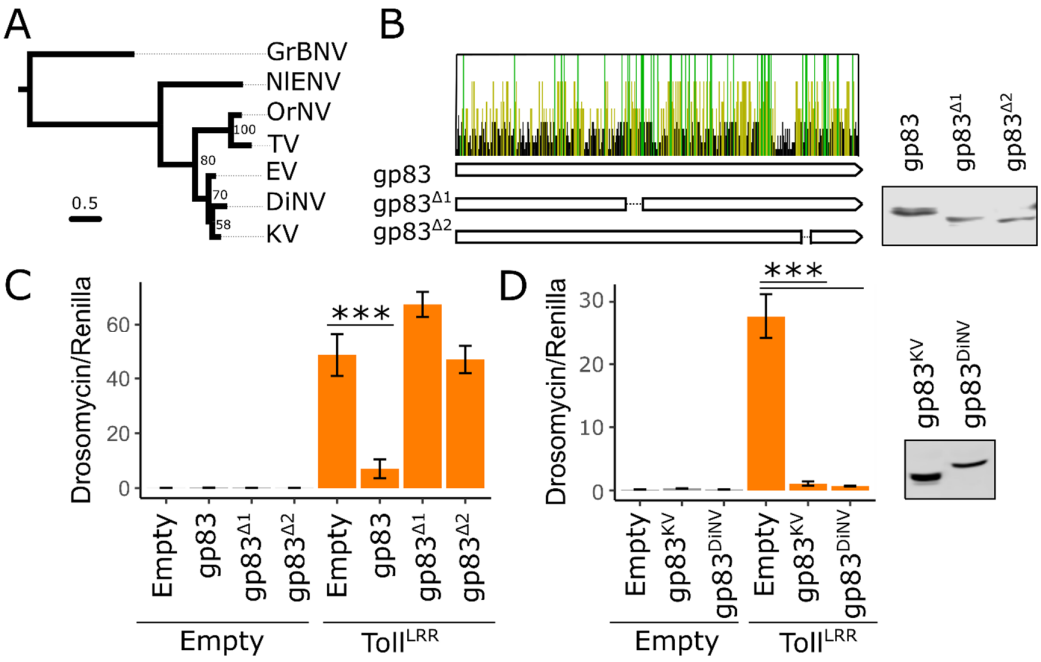

Figure 5: The immunosuppressive function of gp83 is conserved

(A) Conservation of gp83 amino acid residues across 6 species of nudivirus. Each bar represents an amino-acid residue, and bars are coloured yellow if the residue is conserved in ≥ 50% of the species shown in the phylogeny, and green if conserved in 100% of the species. We created two deletion constructs spanning regions with an excess of conserved residues: gp83$^{\Delta1}$ and gp83$^{\Delta2}$. Although protein accumulates to normal levels in both deletion constructs following transfection (A), overexpression of gp83$^{\Delta1}$ or gp83$^{\Delta2}$ failed to inhibit Toll^{LRR}-induced Drs luciferase, indicating these conserved residues are important for its immunosuppressive function. (C) Additionally, gp83 from Drosophila innubila nudivirus (DiNV) was able to inhibit Toll signalling in D. melanogaster cells, indicating the Toll suppressive function against D. melanogaster is conserved in other nudiviruses. Error bars show standard error of the mean (n = 5).

5.4.7 Conclusions

In this study, we investigated the role of known anti-RNA viral immune pathways in the context of DNA virus infection, including RNAi, JAK-STAT, Imd, and Toll pathways. Our data support an antiviral role for RNAi and Imd against KV, consistent with previously-described antiviral RNAi against IIV6 and DNA virus-encoded suppressors of Imd (Bronkhorst *et al.*, 2012;

Kemp *et al.*, 2013; Lamiable, Kellenberger, *et al.*, 2016). Further, we identified gp83 as a KV gene that inhibits Toll signalling downstream of the release of NF-κB transcription factors from IκB, strongly suggesting Toll signalling can be antiviral during DNA virus infection in insects. The immunosuppressive effect of gp83 is conserved in other nudiviruses, and has not evolved host-specificity in DiNV, indicating the Toll-gp83 interaction is unlikely to be a hotspot of reciprocal host-virus adaptation and that other KV genes may be more important in determining host range.

Chapter 6: General discussion

I am the sole author of the following text, with minor comments on earlier drafts from Darren Obbard.

6.1 Summary of the field

Viruses must evade or inhibit host antiviral immune mechanisms and hijack host factors to promote their replication and transmission. The virus and host genes that underlie these pro- and antiviral interactions are therefore in conflict, and may enter an evolutionary arms race, where selection on the host favours resistance alleles and selection on the virus favours strains that overcome or evade host resistance. Because the virus-host interactions that are engaged in an arms race underlie viral resistance (by definition), they contribute to variation in susceptibility to infection within species and to the host range of a virus across species. This is supported by the observation that hosts harbour more genetic variation in resistance to naturally occurring viruses than to non-native viruses (Magwire *et al.*, 2012), and by the tendency of viral-encoded immune inhibitors to be highly species-specific (Parisien, Lau and Horvath, 2002; Mariani *et al.*, 2003; Goffinet *et al.*, 2009; Rajsbaum *et al.*, 2012; van Mierlo *et al.*, 2014; Stabell *et al.*, 2018). Therefore, the identification of co-evolving virus-host interactions, and the description of their evolutionary dynamics, is of general importance to understanding the molecular determinants of virus resistance and host range.

Drosophila melanogaster is an excellent model for addressing these questions, due to its relatively simple demographic history, large population sizes and associated effective selection, and genetic tools allowing dissection of antiviral immune processes. Previous work has described general antiviral mechanisms, such as RNA interference and apoptosis, and virus-specific defences, including autophagy and transcriptional responses orchestrated by NF-KB and JAK-STAT signalling pathways (Kemp *et al.*, 2013; Lamiable, Arnold, *et al.*, 2016). Segregating polymorphism that underlies resistance to native *Drosophila* viruses have also been mapped, and some of these resistance loci have signatures of host-virus coevolution (Magwire *et al.*, 2011, 2012; Martins *et al.*, 2014; Cogni *et al.*, 2016; Cao *et al.*, 2017). Furthermore, recent metagenomic sequencing has identified approximately 100 viruses associated with *Drosophila* species (Webster *et al.*, 2015, 2016; Medd *et al.*, 2018). Although most of these await isolation, the potential of *Drosophila* as a model to study comparative host-virus interactions and their evolution is apparent, and future work should focus on using

comparative approaches to determine the diversity of pro- and antiviral pathways in insects, and associated evolutionary patterns.

6.2 Overview of thesis

6.2.1 Host-virus coevolution in the RNA interference pathway

RNA interference is one of the most general antiviral immune response in insects, and controls infection of disparate viral families (Bronkhorst and van Rij, 2014). The effectiveness of RNAi is mirrored in the evolution of viral suppressors of RNAi (VSRs), which are encoded by the majority of obligate insect viruses studied in the lab (Li, Li and Ding, 2002; Van Rij *et al.*, 2006; Nayak *et al.*, 2010; Joël T van Mierlo *et al.*, 2012; Alfred W Bronkhorst *et al.*, 2014; Van Cleef *et al.*, 2014; van Mierlo *et al.*, 2014). The generality of RNAi, and the resulting diversity of VSRs, make RNAi a likely target of strong pathogen-mediated selection. This hypothesis has been tested in *Drosophila*, where antiviral RNAi genes exhibit genomic patterns consistent with antagonistic coevolution, including increased diversity (Chapter 2, below) and divergence (Chapter 3, below) (Obbard *et al.*, 2006, 2011; Obbard et al., 2009; Darren J. Obbard *et al.*, 2009; Kolaczkowski, Hupalo and Kern, 2011).

In Chapter 2, I investigated the function of the Ago2 N-terminal glutamine-rich repeat region (GRR) in the context of pathogen-mediated selection. The GRR is hypervariable within and across *Drosophila* species, possibly due to diversifying selection, a predicted outcome of some coevolutionary models (Anderson and May, 1982; Antonovics and Thrall, 1994; Sasaki, 2000; Ebert, 2008; Hain *et al.*, 2010; Brockhurst *et al.*, 2014). Through long-read amplicon sequencing I found that there are two divergent haplotypes which are present at approximately equal frequencies in *Drosophila* populations, and that one of these haplotypes has recently risen to high frequency, consistent with either positive selection or a population bottleneck and subsequent expansion in American populations. Further, I found this region likely has a functional role in RNAi, evidenced by conservation of GRR amino acid composition across insect species. However, previous studies have removed the GRR from *Drosophila* Ago2, and observe no significant changes to RISC loading (Liu *et al.*, 2009), hinting at a role specifically in antiviral defence. While I did not find a reproducible association between GRR length or sequence variants and resistance to Drosophila C Virus, GRR variation could be involved in resistance to other *Drosophila* viruses, and future studies should test whether these

divergent haplotypes provide differential protection against other *Drosophila* viruses as more isolates become available.

In Chapter 3, I used population genomic data from 8 invertebrate species and investigated rates of adaptive evolution in RNAi genes, both to extend the observations from *Drosophila* to other a wider taxonomic range, and to gain better estimates for the adaptive rate associated with RNAi genes and RNAi functional classes. I confirmed the elevated adaptive rate observed in *Drosophila* is a general phenomenon in invertebrates, and quantified rates of evolution associated with the miRNA, siRNA, piRNA, and viRNA pathways. I found that the viRNA pathway and piRNA pathways evolve rapidly, most likely due to conflict with viruses and transposable elements, respectively. Further, I used a cross-species linear model-based approach to identify genes with elevated adaptation across the analysed insect species. I identified Ago2, Dcr2, and Hen1 (among other piRNA pathway genes), as having significantly elevated rates of adaptive evolution as compared to a set of control genes, each of which is a known target of certain VSRs (Van Rij *et al.*, 2006; Vogler *et al.*, 2007; Nayak *et al.*, 2010; Joël T. van Mierlo *et al.*, 2012; Van Cleef *et al.*, 2014).

6.2.2 Kallithea virus infection in Drosophila melanogaster

Despite the striking patterns of polymorphism and divergence associated with RNAi genes, there has yet to be definitive evidence that the observed evolution in the RNAi pathway is virus-mediated. Instead, an understanding of virus-mediated selection in *Drosophila* comes primarily from the identification of loci segregating alleles that provide viral resistance (Magwire *et al.*, 2011, 2012, Cao *et al.*, 2016, 2017; Cogni *et al.*, 2016). Under the assumptions that the virus is detrimental to host fitness, and that resistance alleles provide a similar benefit in nature as in the laboratory (Wilfert and Jiggins, 2010), then these loci should be under virus-driven selection. However, because resistance alleles are expected to arise from the coevolutionary process, this requires natural virus-host associations, and there are very few *Drosophila* RNA virus isolates available, and no DNA virus isolates.

In Chapter 4, I isolated Kallithea virus (KV), a *D. melanogaster* DNA virus, and characterised basic host fitness costs associated with infection, virus resistance polymorphisms, and host transcriptional responses to infection. I found that KV affected fitness associated traits in the lab, reducing female fecundity and causing early male-specific lethality. I performed a genome-wide association study by measuring KV titre and KV-induced mortality in the *Drosophila* Genetic Reference Panel (Mackay *et al.*, 2012), and although I found few convincing viral

resistance variants (i.e. reducing virus replication), I identified a nonsynonymous polymorphism in *Cip4* associated with longevity following infection. Finally, I analysed the transcriptional response to infection, and found that many of the highly upregulated genes do not have a previously described role in viral pathogenesis, suggesting a subset of KV-specific or DNA virus-specific responses.

6.2.3 Conflict between NF-κB and Kallithea virus

Anti-DNA viral immunity has not been well studied in *Drosophila*, in part because no DNA virus isolates of *Drosophila* have been available. Instead, Insect Iridescent Virus 6 (IIV6), a moth iridovirus which can infect *Drosophila*, has been used as a model of DNA virus infection. RNAi retains antiviral activity against IIV6, and IIV6 encodes a VSR, extending RNAi-VSR conflict to DNA viruses (Bronkhorst *et al.*, 2012; Bronkhorst *et al.*, 2014). Additionally, apoptosis is a generally effective immune response to baculovirus infection, consistent with the presence of baculovirus-encoded inhibitors of apoptosis (Bump *et al.*, 1995; Xue and Robert Horvitz, 1995; Clem, 2001; Byers, Vandergaast and Friesen, 2016). However, other immune pathways with known anti-RNA viral roles, such as the Toll and Imd pathways, have not been explored in the context of insect DNA virus infections.

In Chapter 5, I presented evidence supporting an antiviral role for RNAi, Imd, and Toll pathways during KV infection. I found that KV infection induced Imd signalling, and that Imd mutants sustained higher viral titre. While I did not observe higher KV titre in Toll mutants, I mapped gp83 as a KV-encoded gene that is able to potently inhibit Toll signalling downstream of NF-KB transcription factors, strongly suggesting Toll is antiviral but efficiently suppressed by KV. Further, this immunosuppressive function is conserved in gp83 from Drosophila innubila nudivirus, which can efficiently silence Toll signalling in *D. melanogaster* cells. Therefore, while there is conflict between nudivirus-encoded gp83 and the *Drosophila* Toll pathway, this interaction has not led to an arms race-like scenario, whereby DiNV gp83 specialises on the *D. innubila* immune system to exclusion of functionality in *D. melanogaster*.

6.3 Future directions

6.3.1 Host-virus coevolution in Drosophila

The genomic patterns associated with antiviral RNAi genes are striking, and it seems likely that this is mediated at least in part by virus-encoded immune suppressors. However, because antiviral RNAi pathway genes also process TE-derived dsRNA, positive selection derived from a function in TE defence cannot be ruled out. Further, these selective pressures are difficult to separate through comparative approaches because of the ubiquity of TEs and viruses, and our incomplete knowledge of their histories in different species. Instead, confirmation of virus-driven RNAi evolution could come from identification of virus resistance variants in antiviral RNAi (Lambrechts *et al.*, 2012), and subsequently following their trajectory through populations. However, assuming a combined α of Dcr2 and Ago2 is 0.64 (the average across 6 insect species; Chapter 3), and 353 combined nonsynonymous changes between *D. melanogaster* and *D. simulans* divergence across 3 mya (Obbard *et al.*, 2012), then an adaptive mutation in Dcr2 or Ago2 will sweep once every 13,200 years. Although this excludes changes in the Ago2 GRR and noncoding regions, it is still unlikely to be witnessed first-hand. Indirect evidence for a VSR-RNAi arms race may also come from species-specificity of VSRs in viruses which have co-evolved with their hosts, as has been shown for *Drosophila* Nora virus suppressors of RNAi (van Mierlo *et al.*, 2014). However, this requires groups of hosts and viruses that have co-diverged, such as the Drosophila Nora viruses (van Mierlo *et al.*, 2014), Drosophila sigma viruses, or possibly Drosophila-associated nudiviruses.

Even if completely virus-driven, positive selection in the RNAi pathway is almost certainly in response to the combined selective pressures of multiple viruses, and likely makes only a small contribution to the host adaptations in response to any single virus. A more holistic view of the virus-host interactions with likely coevolutionary importance can be obtained through mapping of resistance alleles, which has been performed during infections with *D. melanogaster* sigma virus, Drosophila C Virus, and Kallithea virus (Chapter 5) (Magwire *et al.*, 2011, 2012, Cao *et al.*, 2016, 2017; Cogni *et al.*, 2016). Thanks to large genotyped and inbred fly strain panels such as the DSPR and DGRP (King, Macdonald and Long, 2012; Mackay *et al.*, 2012), identifying existing resistance variants in *Drosophila* is relatively fool proof, and currently the limiting factor is a lack of natural *Drosophila* virus isolates. Additionally, a reverse genetics approach could be taken, and genome-wide discovery of the host genes important

during infection could be combined with signatures of antagonistic coevolution to identify candidate arms race genes, a prospect that has recently become much more feasible with the advent of pooled genome-wide CRISPR screening in *Drosophila* cells (Viswanatha *et al.*, 2018). Regardless of approach, the continued identification of the virus-host interactions that underlie the coevolutionary process will allow for broader generalizations of their characteristics, such as whether resistance mutations are enriched in certain stages of the virus replication cycle, in pro- or antiviral host genes, or in conserved or virus-specific interactions.

6.3.2 Nudivirus-Drosophila model

The nudivirus-*Drosophila* model has promise for characterising anti-DNA viral immunity, investigating variation in resistance and tolerance to infection, and understanding the evolution of host specialization.

Insect anti-DNA viral immunity is not well understood, and while anti-RNA viral immune pathways seem to retain a function during DNA virus infection (Bronkhorst *et al.*, 2012; Kemp *et al.*, 2013; West and Silverman, 2018) (Chapter 5), there are likely DNA virus-specific responses that are undescribed. This is reflected in the transcriptional response to infection, where the most highly induced genes have either poorly annotated functions, or lack known protein domains altogether (Chapter 4). Additionally, while anti-RNA viral immunity is primarily mediated by RNAi, both Kallithea virus and IIV6 infections have only modestly increased titre in RNAi mutants (Bronkhorst *et al.*, 2012)(Chapter 5). In the case of IIV6, this may be due to a VSR; however, this cannot explain the similar KV titres in RNAi mutants during later infection stages, as we could not detect KV-mediated RNAi inhibition (Chapter 5). It is therefore possible that the strategy of RNAi changes during DNA virus infection, and primarily mediates a tolerance response through silencing of DNA virus genes rather than reducing KV titre, either transcriptionally or post-transcriptionally. Alternatively, the relative importance of antiviral pathways may be different between RNA and DNA viruses, and KV may be more sensitive to transcriptional or cellular responses to infection. These antiviral immune responses important during KV infection may be determined by a more comprehensive screen of the effect of KV genes on host processes (Chapter 5), both because the increased coding capacity of DNA viruses may allow more modular gene function, and well-studied DNA viruses dedicate a large proportion of their genes to immune interference (Haga and Bowie, 2005; Melchjorsen, Matikainen and Paludan, 2009).

KV infection is associated with multiple phenotypes that likely affect host fitness, such as early male lethality, reduced movement, and a cessation of oogenesis, and it may be expected that these fitness-associated traits harbour substantial genetic variation (Houle, 1992). I quantified genetic variation in the control of viral titre and virus-induced mortality, and surprisingly found relatively low genetic variation in KV titre and few associated variants (as compared to other *Drosophila* viruses). While there may truly be few polymorphisms with large effects on KV titre, the DGRP are underpowered to detect low or high frequency polymorphisms, and it is therefore possible that KV resistance mutations exist in these frequency classes - a hypothesis that could be tested using recombinant inbred lines. Virus-induced mortality had substantially higher amounts of genetic variation, at least partially accounted for by a nonsynonymous SNP in *Cip4*. However, it is unknown whether mortality following infection is associated with other KV-induced traits – that is, whether reduced movement, loss of fecundity, and eventual mortality are all part of an overarching tolerance response affecting fly health. Infection of the DGRP and subsequent monitoring of movement and fecundity would provide genetic correlations between these phenotypes, and may also identify additional or overlapping causal variants, as some flies in a DGRP outbred population were found to escape oogenesis shut down (Chapter 4). Finally, it is still unclear whether these results are generalizable across KV genotypes, and given the prevalence of KV in the wild, additional KV isolates could be used to search for genotype-by-genotype interactions in the control of KV replication and fitness-associated traits following infection.

There are five identified *Drosophila*-associated nudiviruses (Unckless, 2011; Webster *et al.*, 2015; Kapun *et al.*, 2018), raising the possibility of using nudiviruses to study the virus-host interactions that underlie the evolution of species-specificity and cross resistance properties of host variants affecting resistance and tolerance to nudiviruses. For example, KV has evolved some degree of host specificity, and while it has been found associated with *D. melanogaster* and *D. simulans* (Webster *et al.*, 2015), it cannot infect *D. suzukii* (Nathan Medd, personal communication), and thus may be unlikely to infect the more distantly related *D. innubila*. This implies that coevolution of KV with *D. melanogaster* has led to the specialization of some virus-host interactions, and these may be determined by complementing KV infection with combinations of DiNV genes in cells that DiNV can infect, but KV cannot. Additionally, three nudiviruses (Esparto virus, Tomelloso virus, and Mauternbach virus) are asso-

ciated with *D. melanogaster*, and GWAS could be repeated during infection with these nudiviruses to determine genetic correlations in resistance and tolerance to similar viruses, and whether virus resistance variants have cross-resistance properties.

6.3.3 Comparative immunology in Drosophila

Recent work has compared antiviral immune responses across disparate insect-infecting viruses using *Drosophila*, finding both general and virus-specific immune mechanisms (Kemp *et al.*, 2013; Ferreira *et al.*, 2014; Lamiable and Imler, 2014; Lamiable, Arnold, *et al.*, 2016). Additionally, metagenomic sequencing has identified over 100 *Drosophila*-associated viruses, including those from at least 15 families previously unstudied in *Drosophila* (Unckless, 2011; van Mierlo *et al.*, 2014; Longdon, Murray, *et al.*, 2015; Webster *et al.*, 2015, 2016; Kapun *et al.*, 2018; Medd *et al.*, 2018). Continued isolation of *Drosophila*-associated viruses and characterisation of the relative importance of various antiviral immune pathways will lead to a better understanding of the diversity of immune responses, the attributes of infection with which they are associated, and how these immune responses are distributed across the virus phylogeny. Additionally, *Drosophila*-associated viruses include those in families of medical importance, including *Reoviridae*, *Bunyaviridae*, and *Flaviviridae*, suggesting their potential use as models for studying vector-borne disease transmission to vertebrates.

In addition to expanding the diversity of known immune responses, there are significant gaps in our understanding of the currently identified innate immune pathways. While many of the broad pathways implicated in defence have been identified, and transcriptional responses to diverse viruses have been sequenced, the mechanisms of virus recognition, signal integration, and non-RNAi antiviral effectors are poorly understood. For example, JAK-STAT upregulates *vir-1* and is antiviral against dicistroviruses, but activates *TotM* and is not antiviral during nodavirus, alphavirus, birnavirus, or rhabdovirus infections (Dostert *et al.*, 2005; Kemp *et al.*, 2013). Therefore, future research could focus on the identification of viral PAMPs or infection by-products (e.g. dsRNA, defective viral particles, reactive oxygen species, host damage (Deddouche *et al.*, 2008; Poirier *et al.*, 2018; West and Silverman, 2018)) that ultimately signal through transcriptional pathways like JAK-STAT, the host factors which recognise these infection cues, and how these signals converge on a unique transcriptional response. A specific route of interest may be the dissection of the *Drosophila* cGAS-STING pathway, as the *Drosophila* STING orthologue is induced following KV infection, and this pathway is involved

in viral DNA sensing and activation of an innate immune response in mammals (Chen, Sun and Chen, 2016).

6.3.4 Concluding remarks

I have found that the RNA interference pathway exhibits signatures consistent with host-virus coevolution across insect species, including high rates of adaptive protein evolution and dynamic rearrangements of the glutamine rich repeat region of Argonaute-2. I isolated Kallithea virus to expand the repertoire of viruses that can be used for the study of host-virus coevolution in *Drosophila*, and identified host polymorphisms affecting resistance and tolerance, and transcriptional immune responses to infection. Finally, I described interactions between the NF-κB pathways and Kallithea virus, including antiviral NF-κB signalling, and a Kallithea virus-encoded immune inhibitor. These results suggest conflict between innate immune pathways and viruses is widespread, and can be explored through isolation and characterisation of new *Drosophila* viruses.

References

Agnello, V. *et al.* (1999) 'Hepatitis C virus and other flaviviridae viruses enter cells via low density lipoprotein receptor.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 96(22), pp. 12766–71. doi: 10.1073/PNAS.96.22.12766.

Albecka, A. *et al.* (2012) 'Role of low-density lipoprotein receptor in the hepatitis C virus life cycle', *Hepatology*, 55(4), pp. 998–1007. doi: 10.1002/hep.25501.

Alvarez-Garcia, I. and Miska, E. A. (2005) 'MicroRNA functions in animal development and human disease.', *Development (Cambridge, England)*, 132(21), pp. 4653–62. doi: 10.1242/dev.02073.

Ambrose, R. L. *et al.* (2009) 'Drosophila A virus is an unusual RNA virus with a T=3 icosahedral core and permuted RNA-dependent RNA polymerase', *Journal of General Virology*, 90(9), pp. 2191–2200. doi: 10.1099/vir.0.012104-0.

Anders, S. and Huber, W. (2010) 'Differential expression analysis for sequence count data', *Genome Biology*, 11(10), p. R106. doi: 10.1186/gb-2010-11-10-r106.

Anderson, K. V and Nüsslein-Volhard, C. (1984) 'Information for the dorsal-ventral pattern of the Drosophila embryo is stored as maternal mRNA.', *Nature*, 311(5983), pp. 223–7.

Anderson, R. M. and May, R. M. (1982) 'Coevolution of Hosts and Parasites', *Parasitology*, 85(2), pp. 411–426. doi: 10.1017/S0031182000055360.

Antonov, A. V. *et al.* (2009) 'PPI spider: A tool for the interpretation of proteomics data in the context of protein-protein interaction networks', *PROTEOMICS*, 9(10), pp. 2740–2749. doi: 10.1002/pmic.200800612.

Antonov, A. V (2011) 'BioProfiling.de: analytical web portal for high-throughput cell biology.', *Nucleic acids research*. Oxford University Press, 39(Web Server issue), pp. W323-7. doi: 10.1093/nar/gkr372.

Antonovics, J. and Thrall, P. H. (1994) 'The Cost of Resistance and the Maintenance of Genetic Polymorphism in Host-Pathogen Systems', *Proceedings: Biological Sciences*. The Royal Society, 257(1349), pp. 105–110.

Aravin, A. A., Hannon, G. J. and Brennecke, J. (2007) 'The Piwi-piRNA Pathway Provides an Adaptive Defense in the Transposon Arms Race', *Science*, 318(5851), pp. 761–764. doi: 10.1126/science.1146484.

Arbouzova, N. I. and Zeidler, M. P. (2006) 'JAK/STAT signalling in Drosophila: insights into conserved regulatory and cellular functions.', *Development (Cambridge, England)*, 133(14), pp. 2605–16. doi: 10.1242/dev.02411.

Avadhanula, V. *et al.* (2009) 'A Novel System for the Launch of Alphavirus RNA Synthesis Reveals a Role for the Imd Pathway in Arthropod Antiviral Response', *PLoS Pathogens*. Edited by S.-W. Ding. Public Library of Science, 5(9), p. e1000582. doi: 10.1371/journal.ppat.1000582.

Avila, A. *et al.* (2002) 'The Drosophila Atypical Protein Kinase C-Ref(2)P Complex Constitutes a Conserved Module for Signaling in the Toll Pathway', *Molecular and Cellular Biology*. American Society for Microbiology, 22(24), pp. 8787–8795. doi: 10.1128/MCB.22.24.8787-

8795.2002.

Ayres, J. S. and Schneider, D. S. (2009) 'The Role of Anorexia in Resistance and Tolerance to Infections in Drosophila', *PLoS Biology*. Edited by D. Promislow. Public Library of Science, 7(7), p. e1000150. doi: 10.1371/journal.pbio.1000150.

Baeg, G.-H., Zhou, R. and Perrimon, N. (2005) 'Genome-wide RNAi analysis of JAK/STAT signaling components in Drosophila', *Genes & Development*, 19(16), pp. 1861–1870. doi: 10.1101/gad.1320705.

Bangham, J. *et al.* (2007) 'The age and evolution of an antiviral resistance mutation in Drosophila melanogaster.', *Proceedings. Biological sciences / The Royal Society*, 274(1621), pp. 2027–2034. doi: 10.1098/rspb.2007.0611.

Bangham, J. *et al.* (2008) 'Genetic variation affecting host-parasite interactions: Major-effect quantitative trait loci affect the transmission of sigma virus in Drosophila melanogaster', *Molecular Ecology*, 17(17), pp. 3800–3807. doi: 10.1111/j.1365-294X.2008.03873.x.

Bankevich, A. *et al.* (2012) 'SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing', *Journal of Computational Biology*, 19(5), pp. 455–477. doi: 10.1089/cmb.2012.0021.

Barrière, A. and Félix, M.-A. (2005) 'High local genetic diversity and low outcrossing rate in Caenorhabditis elegans natural populations.', *Current biology : CB*, 15(13), pp. 1176–84. doi: 10.1016/j.cub.2005.06.022.

Barton, N. H. (1998) 'The effect of hitch-hiking on neutral genealogies', *Genetical Research*, 72(2), p. S0016672398003462. doi: 10.1017/S0016672398003462.

Benson, G. (1999) 'Tandem repeats finder: a program to analyze DNA sequences.', *Nucleic acids research*, 27(2), pp. 573–80.

Beye, M. *et al.* (2006) 'Exceptionally high levels of recombination across the honey bee genome.', *Genome research*, 16(11), pp. 1339–44. doi: 10.1101/gr.5680406.

Bierne, N. and Eyre-Walker, A. (2004) 'The genomic rate of adaptive amino acid substitution in Drosophila.', *Molecular biology and evolution*, 21(7), pp. 1350–60. doi: 10.1093/molbev/msh134.

Bitra, K., Suderman, R. J. and Strand, M. R. (2012) 'Polydnavirus Ank proteins bind NF-κB homodimers and inhibit processing of Relish.', *PLoS pathogens*. Edited by D. S. Schneider, 8(5), p. e1002722. doi: 10.1371/journal.ppat.1002722.

Blissard, G. W. and Rohrmann, G. F. (1990) 'Baculovirus Diversity and Molecular Biology', *Annual Review of Entomology*. Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA , 35(1), pp. 127–155. doi: 10.1146/annurev.en.35.010190.001015.

Blumenstiel, J. P., Erwin, A. A. and Hemmer, L. W. (2016) 'What Drives Positive Selection in the Drosophila piRNA Machinery? The Genomic Autoimmunity Hypothesis.', *The Yale journal of biology and medicine*. Yale Journal of Biology and Medicine, 89(4), pp. 499–512.

Boland, A. *et al.* (2010) 'Crystal structure and ligand binding of the MID domain of a eukaryotic Argonaute protein.', *EMBO reports*, 11(7), pp. 522–7. doi:

10.1038/embor.2010.81.

Bou Sleiman, M. S. *et al.* (2015) 'Genetic, molecular and physiological basis of variation in Drosophila gut immunocompetence.', *Nature communications*, 6, p. 7829. doi: 10.1038/ncomms8829.

Boucias, D. G., Maruniak, J. E. and Pendland, J. C. (1989) 'Characterization of a non-occluded baculovirus (subgroup C) from the field cricket,Gryllus rubens', *Archives of Virology*. Springer-Verlag, 106(1–2), pp. 93–102. doi: 10.1007/BF01311041.

Breitenbach, J. E., Shelby, K. S. and Popham, H. J. R. (2011) 'Baculovirus Induced Transcripts in Hemocytes from the Larvae of Heliothis virescens', *Viruses*. Molecular Diversity Preservation International, 3(12), pp. 2047–2064. doi: 10.3390/v3112047.

Brennecke, J. *et al.* (2007) 'Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in Drosophila', *Cell*, 128(6), pp. 1089–1103. doi: 10.1016/j.cell.2007.01.043.

Brockhurst, M. A. *et al.* (2014) 'Running with the Red Queen: the role of biotic conflicts in evolution', *Proceedings of the Royal Society of London B: Biological Sciences*, 281(1797).

Bronkhorst, A. W. *et al.* (2012) 'The DNA virus Invertebrate iridescent virus 6 is a target of the Drosophila RNAi machinery', *Proceedings of the National Academy of Sciences*, 109(51), pp. E3604–E3613. doi: 10.1073/pnas.1207213109.

Bronkhorst, A. W. *et al.* (2014) 'A dsRNA-binding protein of a complex invertebrate DNA virus suppresses the Drosophila RNAi response.', *Nucleic acids research*. Oxford University Press, 42(19), pp. 12237–48. doi: 10.1093/nar/gku910.

Bronkhorst, A. W. *et al.* (2014) 'A dsRNA-binding protein of a complex invertebrate DNA virus suppresses the Drosophila RNAi response', *Nucleic Acids Research*, 42(19), pp. 12237–12248. doi: 10.1093/nar/gku910.

Bronkhorst, A. W. and van Rij, R. P. (2014) 'The long and short of antiviral defense: small RNA-based immunity in insects', *Current Opinion in Virology*, 7, pp. 19–28. doi: 10.1016/j.coviro.2014.03.010.

Brun, P. and Plus, N. (1980) 'The viruses of Drosophila', in *The genetics and biology of Drosophila*, pp. 625–702.

Buckling, A. and Rainey, P. B. (2002) 'Antagonistic coevolution between a bacterium and a bacteriophage.', *Proceedings. Biological sciences*. The Royal Society, 269(1494), pp. 931–6. doi: 10.1098/rspb.2001.1945.

Bump, N. J. *et al.* (1995) 'Inhibition of ICE family proteases by baculovirus antiapoptotic protein p35.', *Science (New York, N.Y.)*, 269(5232), pp. 1885–8.

Burand, J. P. *et al.* (2012) 'Analysis of the Genome of the Sexually Transmitted Insect Virus Helicoverpa zea Nudivirus 2', *Viruses*. Molecular Diversity Preservation International, 4(12), pp. 28–61. doi: 10.3390/v4010028.

Burand, J. P., Stiles, B. and Wood, H. A. (1983) 'Structural and Intracellular Proteins of the Nonoccluded Baculovirus HZ-1.', *Journal of virology*, 46(1), pp. 137–42.

Byers, N. M., Vandergaast, R. L. and Friesen, P. D. (2016) 'Baculovirus Inhibitor-of-Apoptosis Op-IAP3 Blocks Apoptosis by Interaction with and Stabilization of a Host Insect Cellular IAP',

Journal of Virology. Edited by G. McFadden, 90(1), pp. 533–544. doi: 10.1128/JVI.02320-15.

Camacho, C. *et al.* (2009) 'BLAST+: architecture and applications.', *BMC bioinformatics*. BioMed Central, 10(1), p. 421. doi: 10.1186/1471-2105-10-421.

Cao, C. *et al.* (2016) 'A Polymorphism in the Processing Body Component Ge-1 Controls Resistance to a Naturally Occurring Rhabdovirus in Drosophila', *PLoS Pathogens*. Edited by D. S. Schneider, 12(1), p. e1005387. doi: 10.1371/journal.ppat.1005387.

Cao, C. *et al.* (2017) 'Complex coding and regulatory polymorphisms in a restriction factor determine the susceptibility of Drosophila to viral infection', *Genetics*, 206(4), pp. 2159–2173. doi: 10.1534/genetics.117.201970.

Carmell, M. A. *et al.* (2002) 'The Argonaute family: tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis.', *Genes & development*, 16(21), pp. 2733–42. doi: 10.1101/gad.1026102.

Carpenter, J. A. *et al.* (2009) 'Evidence for ADAR-induced hypermutation of the Drosophila sigma virus (Rhabdoviridae)', *BMC Genetics*. BioMed Central, 10, p. 75. doi: 10.1186/1471-2156-10-75.

Carré-Mlouka, A. *et al.* (2007) 'Control of sigma virus multiplication by the ref(2)P gene of Drosophila melanogaster: An in vivo study of the PB1 domain of Ref(2)P', *Genetics*. Genetics Society of America, 176(1), pp. 409–419. doi: 10.1534/genetics.106.063826.

Charlesworth, B. (1994) 'The effect of background selection against deleterious mutations on weakly selected, linked variants.', *Genetical research*, 63(3), pp. 213–27.

Charlesworth, B. (2009) 'Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation.', *Nature reviews. Genetics*. Nature Publishing Group, 10(3), pp. 195–205. doi: 10.1038/nrg2526.

Charlesworth, B., Sniegowski, P. and Stephan, W. (1994) 'The evolutionary dynamics of repetitive DNA in eukaryotes.', *Nature*, 371(6494), pp. 215–20. doi: 10.1038/371215a0.

Charlesworth, J. and Eyre-Walker, A. (2008) 'The McDonald-Kreitman Test and Slightly Deleterious Mutations', *Molecular Biology and Evolution*, 25(6), pp. 1007–1015. doi: 10.1093/molbev/msn005.

Chen, Q., Sun, L. and Chen, Z. J. (2016) 'Regulation and function of the cGAS–STING pathway of cytosolic DNA sensing', *Nature Immunology*. Nature Research, 17(10), pp. 1142–1149. doi: 10.1038/ni.3558.

Chen, Y.-W. *et al.* (2014) 'Systematic study of Drosophila microRNA functions using a collection of targeted knockout mutations.', *Developmental cell*, 31(6), pp. 784–800. doi: 10.1016/j.devcel.2014.11.029.

Choi, J. Y. and Aquadro, C. F. (2015) 'Molecular Evolution of Drosophila Germline Stem Cell and Neural Stem Cell Regulating Genes.', *Genome biology and evolution*. Oxford University Press, 7(11), pp. 3097–114. doi: 10.1093/gbe/evv207.

Chuang, H.-C. *et al.* (2011) 'The kinase GLK controls autoimmunity and NF-?B signaling by activating the kinase PKC-? in T cells', *Nature Immunology*. Nature Research, 12(11), pp. 1113–1118. doi: 10.1038/ni.2121.

van Cleef, K. W. R. *et al.* (2011) 'Identification of Viral Suppressors of RNAi by a Reporter

Assay in Drosophila S2 Cell Culture', in *Methods in molecular biology (Clifton, N.J.)*, pp. 201–213. doi: 10.1007/978-1-61779-037-9_12.

Van Cleef, K. W. R. *et al.* (2014) 'Mosquito and Drosophila entomobirnaviruses suppress dsRNA- and siRNA-induced RNAi', *Nucleic Acids Research*. Oxford University Press, 42(13), pp. 8732–8744. doi: 10.1093/nar/gku528.

Clem, R. J. (2001) 'Baculoviruses and apoptosis: the good, the bad, and the ugly', *Cell Death and Differentiation*. Nature Publishing Group, 8(2), pp. 137–143. doi: 10.1038/sj.cdd.4400821.

Cogni, R. *et al.* (2016) 'The genetic architecture of resistance to virus infection in *Drosophila*', *Molecular Ecology*, 25(20), pp. 5228–5241. doi: 10.1111/mec.13769.

Compton, A. A., Hirsch, V. M. and Emerman, M. (2012) 'The host restriction factor APOBEC3G and retroviral Vif protein coevolve due to ongoing genetic conflict.', *Cell host & microbe*, 11(1), pp. 91–8. doi: 10.1016/j.chom.2011.11.010.

Contamine, D., Petitjean, A. M. and Ashburner, M. (1989) 'Genetic resistance to viral infection: the molecular cloning of a Drosophila gene that restricts infection by the rhabdovirus sigma.', *Genetics*, 123(3), pp. 525–33.

Costa, A. *et al.* (2009) 'The Imd pathway is involved in antiviral immune responses in Drosophila', *PLoS ONE*. Edited by P. Sommer. Public Library of Science, 4(10), p. e7436. doi: 10.1371/journal.pone.0007436.

Cox, J. and Mann, M. (2008) 'MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification', *Nature Biotechnology*, 26(12), pp. 1367–1372. doi: 10.1038/nbt.1511.

Crawford, A. M. and Sheehan, C. (1985) 'Replication of Oryctes Baculovirus in Cell Culture: Viral Morphogenesis, Infectivity and Protein Synthesis', *Journal of General Virology*. Microbiology Society, 66(3), pp. 529–539. doi: 10.1099/0022-1317-66-3-529.

Crawford, J. E. and Lazzaro, B. P. (2010) 'The demographic histories of the M and S molecular forms of Anopheles gambiae s.s.', *Molecular biology and evolution*, 27(8), pp. 1739–44. doi: 10.1093/molbev/msq070.

Cutter, A. D. *et al.* (2006) 'Patterns of nucleotide polymorphism distinguish temperate and tropical wild isolates of Caenorhabditis briggsae.', *Genetics*, 173(4), pp. 2021–31. doi: 10.1534/genetics.106.058651.

Czech, B. *et al.* (2008) 'An endogenous small interfering RNA pathway in Drosophila.', *Nature*. Nature Publishing Group, 453(7196), pp. 798–802. doi: 10.1038/nature07007.

Czech, B. *et al.* (2013) 'A transcriptome-wide RNAi screen in the Drosophila ovary reveals factors of the germline piRNA pathway.', *Molecular cell*, 50(5), pp. 749–61. doi: 10.1016/j.molcel.2013.04.007.

Czech, B. *et al.* (2016) 'One Loop to Rule Them All: The Ping-Pong Cycle and piRNA-Guided Silencing', *Trends in Biochemical Sciences*. Elsevier, 41(4), pp. 324–337. doi: 10.1016/j.tibs.2015.12.008.

Das, P. P. *et al.* (2008) 'Piwi and piRNAs act upstream of an endogenous siRNA pathway to suppress Tc3 transposon mobility in the Caenorhabditis elegans germline.', *Molecular cell*,

31(1), pp. 79–90. doi: 10.1016/j.molcel.2008.06.003.

Deddouche, S. *et al.* (2008) 'The DExD/H-box helicase Dicer-2 mediates the induction of antiviral activity in drosophila', *Nature Immunology*, 9(12), pp. 1425–1432. doi: 10.1038/ni.1664.

Delaney, J. R. *et al.* (2006) 'Cooperative control of Drosophila immune responses by the JNK and NF-κB signaling pathways', *The EMBO Journal*. European Molecular Biology Organization, 25(13), pp. 3068–3077. doi: 10.1038/sj.emboj.7601182.

Deng, H. *et al.* (2015) 'Enhanced enteroviral infectivity via viral protease-mediated cleavage of Grb2-associated binder 1.', *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*. The Federation of American Societies for Experimental Biology, 29(11), pp. 4523–31. doi: 10.1096/fj.15-274829.

Deng, H. *et al.* (2017) 'Cleavage of Grb2-Associated Binding Protein 2 by Viral Proteinase 2A during Coxsackievirus Infection.', *Frontiers in cellular and infection microbiology*. Frontiers Media SA, 7, p. 85. doi: 10.3389/fcimb.2017.00085.

Denver, D. R. *et al.* (2009) 'A genome-wide view of Caenorhabditis elegans base-substitution mutation processes.', *Proceedings of the National Academy of Sciences of the United States of America*, 106(38), pp. 16310–4. doi: 10.1073/pnas.0904895106.

DePristo, M. A. *et al.* (2011) 'A framework for variation discovery and genotyping using next-generation DNA sequencing data.', *Nature genetics*, 43(5), pp. 491–8. doi: 10.1038/ng.806.

Deshpande, G., Calhoun, G. and Schedl, P. (2005) 'Drosophila argonaute-2 is required early in embryogenesis for the assembly of centric/centromeric heterochromatin, nuclear division, nuclear migration, and germ-cell formation.', *Genes & development*, 19(14), pp. 1680–5. doi: 10.1101/gad.1316805.

Dobin, A. *et al.* (2013) 'STAR: ultrafast universal RNA-seq aligner', *Bioinformatics*. Oxford University Press, 29(1), pp. 15–21. doi: 10.1093/bioinformatics/bts635.

Dostert, C. *et al.* (2005) 'The Jak-STAT signaling pathway is required but not sufficient for the antiviral response of drosophila', *Nature Immunology*. Nature Publishing Group, 6(9), pp. 946–953. doi: 10.1038/ni1237.

Dru, P. *et al.* (1993) 'Unusual variability of the Drosophila melanogaster ref(2)P protein which controls the multiplication of sigma rhabdovirus', *Genetics*, 133(4), pp. 943–954.

Duchaine, T. F. *et al.* (2006) 'Functional proteomics reveals the biochemical niche of C. elegans DCR-1 in multiple small-RNA-mediated pathways.', *Cell*, 124(2), pp. 343–54. doi: 10.1016/j.cell.2005.11.036.

Dupuis, S. *et al.* (2003) 'Impaired response to interferon-α/β and lethal viral disease in human STAT1 deficiency', *Nature Genetics*, 33(3), pp. 388–391. doi: 10.1038/ng1097.

Ebert, D. (2008) 'Host-parasite coevolution: Insights from the Daphnia-parasite model system', *Current Opinion in Microbiology*, 11(3), pp. 290–301. doi: 10.1016/j.mib.2008.05.012.

Eddy, S. R. (2008) 'A probabilistic model of local sequence alignment that simplifies statistical significance estimation.', *PLoS computational biology*. Public Library of Science,

4(5), p. e1000069. doi: 10.1371/journal.pcbi.1000069.

Edgar, R. C. (2004) 'MUSCLE: multiple sequence alignment with high accuracy and high throughput.', *Nucleic acids research*, 32(5), pp. 1792–7. doi: 10.1093/nar/gkh340.

Eilertson, K. E., Booth, J. G. and Bustamante, C. D. (2012) 'SnIPRE: selection inference using a Poisson random effects model.', *PLoS computational biology*. Public Library of Science, 8(12), p. e1002806. doi: 10.1371/journal.pcbi.1002806.

Elde, N. C. *et al.* (2012) 'Poxviruses deploy genomic accordions to adapt rapidly against host antiviral defenses.', *Cell*. NIH Public Access, 150(4), pp. 831–41. doi: 10.1016/j.cell.2012.05.049.

Enard, D. *et al.* (2016) 'Viruses are a dominant driver of protein adaptation in mammals', *eLife*. eLife Sciences Publications Limited, 5, pp. 56–65. doi: 10.7554/eLife.12469.

Evgen'ev, M. B. *et al.* (1997) 'Penelope, a new family of transposable elements and its possible role in hybrid dysgenesis in Drosophila virilis', *Proceedings of the National Academy of Sciences*, 94(1), pp. 196–201. doi: 10.1073/pnas.94.1.196.

Eyre-Walker, A. (2002) 'Changing effective population size and the McDonald-Kreitman test.', *Genetics*, 162(4), pp. 2017–24.

Eyre-Walker, A. and Keightley, P. D. (2009) 'Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change.', *Molecular biology and evolution*, 26(9), pp. 2097–108. doi: 10.1093/molbev/msp119.

Fan, H.-C. *et al.* (2014) 'Polyglutamine (PolyQ) Diseases: Genetics to Treatments', *Cell Transplantation*. Cognizant Communication Corporation, 23(4), pp. 441–458. doi: 10.3727/096368914X678454.

Fay, J. C., Wyckoff, G. J. and Wu, C.-I. (2001) 'Positive and Negative Selection on the Human Genome', *Genetics*, 158(3), pp. 1227–1234.

Fay, J. C., Wyckoff, G. J. and Wu, C.-I. (2002) 'Testing the neutral theory of molecular evolution with genomic data from Drosophila.', *Nature*, 415(6875), pp. 1024–6. doi: 10.1038/4151024a.

Fay, N. and Panté, N. (2015) 'Nuclear entry of DNA viruses.', *Frontiers in microbiology*. Frontiers Media SA, 6, p. 467. doi: 10.3389/fmicb.2015.00467.

Feng, Z. *et al.* (2013) 'A pathogenic picornavirus acquires an envelope by hijacking cellular membranes.', *Nature*, 496(7445), pp. 367–71. doi: 10.1038/nature12029.

Ferguson, H. M. and Read, A. F. (2002) 'Why is the effect of malaria parasites on mosquito survival still unresolved?', *Trends in Parasitology*, 18(6), pp. 256–261. doi: 10.1016/S1471-4922(02)02281-X.

Ferree, P. M. and Barbash, D. A. (2007) 'Distorted sex ratios: a window into RNAi-mediated silencing.', *PLoS biology*. Public Library of Science, 5(11), p. e303. doi: 10.1371/journal.pbio.0050303.

Ferreira, Á. G. *et al.* (2014) 'The Toll-Dorsal Pathway Is Required for Resistance to Viral Oral Infection in Drosophila', *PLoS Pathogens*. Edited by D. S. Schneider. Public Library of Science, 10(12), p. e1004507. doi: 10.1371/journal.ppat.1004507.

Ferrer-Admetlla, A. *et al.* (2014) 'On detecting incomplete soft or hard selective sweeps using haplotype structure.', *Molecular biology and evolution*, 31(5), pp. 1275–91. doi: 10.1093/molbev/msu077.

Finkelshtein, D. *et al.* (2013) 'LDL receptor and its family members serve as the cellular receptors for vesicular stomatitis virus.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 110(18), pp. 7306–11. doi: 10.1073/pnas.1214441110.

Fiston-Lavier, A.-S. *et al.* (2010) 'Drosophila melanogaster recombination rate calculator.', *Gene*, 463(1–2), pp. 18–20. doi: 10.1016/j.gene.2010.04.015.

Fleuriet, A. (1976) 'PRESENCE OF THE HEREDITARY RHABDOVIRUS SIGMA AND POLYMORPHISM FOR A GENE FOR RESISTANCE TO THIS VIRUS IN NATURAL POPULATIONS OF *DROSOPHILA MELANOGASTER*', *Evolution*, 30(4), pp. 735–739. doi: 10.1111/j.1558-5646.1976.tb00953.x.

Fleuriet, A. (1986) 'Perpetuation of the hereditary sigma virus in populations of its host, Drosophila melanogaster. Geographical analysis of correlated polymorphisms', *Genetica*. Kluwer Academic Publishers, 70(3), pp. 167–177. doi: 10.1007/BF00122183.

Fleuriet, A. and Periquet, G. (1993) 'Evolution of theDrosophila melanogaster-sigma virus system in natural populations from Languedoc (southern France)', *Archives of Virology*. Springer-Verlag, 129(1–4), pp. 131–143. doi: 10.1007/BF01316890.

Fleuriet, A. and Sperlich, D. (1992) 'Evolution of the Drosophila melanogaster-sigma virus system in a natural population from Tuebingen', *Theoretical and Applied Genetics*, 85(2–3), pp. 186–189. doi: 10.1007/BF00222858.

Flores, H. A. *et al.* (2015) 'Adaptive evolution of genes involved in the regulation of germline stem cells in Drosophila melanogaster and D. simulans.', *G3 (Bethesda, Md.)*. Genetics Society of America, 5(4), pp. 583–92. doi: 10.1534/g3.114.015875.

Foley, E. and O'Farrell, P. H. (2004) 'Functional Dissection of an Innate Immune Response by a Genome-Wide RNAi Screen', *PLoS Biology*. Edited by Michael Levine, 2(8), p. e203. doi: 10.1371/journal.pbio.0020203.

Fragkoudis, R. *et al.* (2008) 'Semliki Forest virus strongly reduces mosquito host defence signaling', *Insect Molecular Biology*, 17(6), pp. 647–656. doi: 10.1111/j.1365-2583.2008.00834.x.

Fricke, R. *et al.* (2009) 'Drosophila Cip4/Toca-1 Integrates Membrane Trafficking and Actin Dynamics through WASP and SCAR/WAVE', *Current Biology*, 19(17), pp. 1429–1437. doi: 10.1016/j.cub.2009.07.058.

Gauthier, J., Drezen, J.-M. and Herniou, E. A. (2017) 'The recurrent domestication of viruses: major evolutionary transitions in parasitic wasps.', *Parasitology*, pp. 1–13. doi: 10.1017/S0031182017000725.

Gautier, M. and Vitalis, R. (2012) 'rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure.', *Bioinformatics (Oxford, England)*, 28(8), pp. 1176–7. doi: 10.1093/bioinformatics/bts115.

Gay, P. (1978) '[Drosophila genes which intervene in multiplication of sigma virus (author's transl)]', *Mol Gen Genet*, 159(3), pp. 269–283.

Gemayel, R. *et al.* (2015) 'Variable Glutamine-Rich Repeats Modulate Transcription Factor Activity.', *Molecular cell*, 59(4), pp. 615–27. doi: 10.1016/j.molcel.2015.07.003.

Georgel, P. *et al.* (2001) 'Drosophila Immune Deficiency (IMD) Is a Death Domain Protein that Activates Antibacterial Defense and Can Promote Apoptosis', *Developmental Cell*, 1(4), pp. 503–514. doi: 10.1016/S1534-5807(01)00059-4.

Gerber, H. *et al.* (1994) 'Transcriptional activation modulated by homopolymeric glutamine and proline stretches', *Science*. American Association for the Advancement of Science, 263(5148), pp. 808–811. doi: 10.1126/science.8303297.

Ghildiyal, M. *et al.* (2008) 'Endogenous siRNAs Derived from Transposons and mRNAs in Drosophila Somatic Cells', *Science*, 320(5879), pp. 1077–1081. doi: 10.1126/science.1157396.

Goffinet, C. *et al.* (2009) 'HIV-1 antagonism of CD317 is species specific and involves Vpu-mediated proteasomal degradation of the restriction factor.', *Cell host & microbe*, 5(3), pp. 285–97. doi: 10.1016/j.chom.2009.01.009.

Goic, B. *et al.* (2013) 'RNA-mediated interference and reverse transcription control the persistence of RNA viruses in the insect model Drosophila', *Nature Immunology*. Nature Publishing Group, 14(4), pp. 396–403. doi: 10.1038/ni.2542.

Goic, B. *et al.* (2016) 'Virus-derived DNA drives mosquito vector tolerance to arboviral infection', *Nature Communications*. Nature Publishing Group, 7, p. 12410. doi: 10.1038/ncomms12410.

Gossmann, T. I. *et al.* (2010) 'Genome wide analyses reveal little evidence for adaptive evolution in many plant species.', *Molecular biology and evolution*, 27(8), pp. 1822–32. doi: 10.1093/molbev/msq079.

Gossmann, T. I., Keightley, P. D. and Eyre-Walker, A. (2012) 'The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes.', *Genome biology and evolution*, 4(5), pp. 658–67. doi: 10.1093/gbe/evs027.

Graveley, B. R. *et al.* (2011) 'The developmental transcriptome of Drosophila melanogaster', *Nature*, 471(7339), pp. 473–479. doi: 10.1038/nature09715.

Guan, X. *et al.* (2006) 'Mutation of TweedleD, a member of an unconventional cuticle protein family, alters body shape in Drosophila.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 103(45), pp. 16794–9. doi: 10.1073/pnas.0607616103.

Gupta, V. *et al.* (2017) 'Costs and benefits of sublethal Drosophila C virus infection', *Journal of Evolutionary Biology*, 30(7), pp. 1325–1335. doi: 10.1111/jeb.13096.

Gupta, V. and Vale, P. F. (2017) 'Nonlinear disease tolerance curves reveal distinct components of host responses to viral infection.', *Royal Society open science*, 4(7), p. 170342. doi: 10.1098/rsos.170342.

Ha, M. and Kim, V. N. (2014) 'Regulation of microRNA biogenesis.', *Nature reviews. Molecular cell biology*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 15(8), pp. 509–524. doi: 10.1038/nrm3838.

Habayeb, M. S., Ekström, J. O. and Hultmark, D. (2009) 'Nora virus persistent infections are

not affected by the RNAi machinery', *PLoS ONE*. Edited by M. G. Masucci. Public Library of Science, 4(5), p. e5731. doi: 10.1371/journal.pone.0005731.

Haddrill, P. R., Loewe, L. and Charlesworth, B. (2010) 'Estimating the parameters of selection on nonsynonymous mutations in Drosophila pseudoobscura and D. miranda.', *Genetics*, 185(4), pp. 1381–96. doi: 10.1534/genetics.110.117614.

Hadfield, J. D. (2010) 'MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package', *Journal of Statistical Software*, 33(2), pp. 1–22. doi: 10.18637/jss.v033.i02.

Haga, I. R. and Bowie, A. G. (2005) 'Evasion of innate immunity by vaccinia virus', *Parasitology*, 130(S1), p. S11. doi: 10.1017/S0031182005008127.

Hain, D. *et al.* (2010) 'Natural Variation of the Amino-Terminal Glutamine-Rich Domain in Drosophila Argonaute2 Is Not Associated with Developmental Defects', *PLoS ONE*. Edited by A. Bergmann, 5(12), pp. 1–14. doi: 10.1371/journal.pone.0015264.

Hall, I. M., Noma, K.-I. and Grewal, S. I. S. (2003) 'RNA interference machinery regulates chromosome dynamics during mitosis and meiosis in fission yeast.', *Proceedings of the National Academy of Sciences of the United States of America*, 100(1), pp. 193–8. doi: 10.1073/pnas.232688099.

Hamm, J. J., Carpenter, J. E. and Styer, E. L. (1996) 'Oviposition Day Effect on Incidence of Agonadal Progeny of Helicoverpa zea (Lepidoptera: Noctuidae) Infected with a Virus', *Annals of the Entomological Society of America*. Oxford University Press, 89(2), pp. 266–275. doi: 10.1093/aesa/89.2.266.

Han, B. W. *et al.* (2015) 'piRNA-guided transposon cleavage initiates Zucchini-dependent, phased piRNA production', *Science*, 348(6236), pp. 817–821. doi: 10.1126/science.aaa1264.

Han, M. *et al.* (2013) 'Evolutionary rate patterns of genes involved in the Drosophila Toll and Imd signaling pathway', *BMC Evolutionary Biology*, 13(1), p. 245. doi: 10.1186/1471-2148-13-245.

Handler, D. *et al.* (2013) 'The genetic makeup of the Drosophila piRNA pathway.', *Molecular cell*, 50(5), pp. 762–77. doi: 10.1016/j.molcel.2013.04.031.

Harpur, B. A. *et al.* (2014) 'Population genomics of the honey bee reveals strong signatures of positive selection on worker traits.', *Proceedings of the National Academy of Sciences of the United States of America*, 111(7), pp. 2614–9. doi: 10.1073/pnas.1315506111.

Harris, R. S. (2007) 'Improved pairwise alignment of genomic dna'. Pennsylvania State University.

Hecht, P. M. and Anderson, K. V (1993) 'Genetic characterization of tube and pelle, genes required for signaling between Toll and dorsal in the specification of the dorsal-ventral pattern of the Drosophila embryo.', *Genetics*, 135(2), pp. 405–17.

Hedengren, M. *et al.* (1999) 'Relish, a central factor in the control of humoral but not cellular immunity in Drosophila.', *Molecular cell*, 4(5), pp. 827–37.

Heger, A. and Ponting, C. P. (2007) 'Evolutionary rate analyses of orthologs and paralogs from 12 Drosophila genomes.', *Genome research*, 17(12), pp. 1837–49. doi: 10.1101/gr.6249707.

Herniou, E. A. *et al.* (2004) 'Ancient coevolution of baculoviruses and their insect hosts.', *Journal of virology*. American Society for Microbiology, 78(7), pp. 3244–51. doi: 10.1128/JVI.78.7.3244-3251.2004.

Herniou, E. A. *et al.* (2013) 'When parasitic wasps hijacked viruses: genomic and functional evolution of polydnaviruses.', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*. The Royal Society, 368(1626), p. 20130051. doi: 10.1098/rstb.2013.0051.

Hill, T. and Unckless, R. L. (2017a) 'Baculovirus Molecular Evolution via Gene Turnover and Recurrent Positive Selection of Key Genes', *Journal of Virology*. Edited by J. K. Pfeiffer, 91(22), pp. e01319-17. doi: 10.1128/JVI.01319-17.

Hill, T. and Unckless, R. L. (2017b) 'The dynamic evolution of Drosophila innubila Nudivirus', *Infection, Genetics and Evolution*. doi: 10.1016/j.meegid.2017.11.013.

Hill, T. and Unckless, R. L. (2018) 'The dynamic evolution of Drosophila innubila Nudivirus', *Infection, Genetics and Evolution*. Elsevier, 57, pp. 151–157. doi: 10.1016/J.MEEGID.2017.11.013.

Hill, W. G. and Robertson, A. (1968) 'Linkage disequilibrium in finite populations.', *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, 38(6), pp. 226–31. doi: 10.1007/BF01245622.

Hopkins, K. C. *et al.* (2013) 'A genome-wide RNAi screen reveals that mRNA decapping restricts bunyaviral replication by limiting the pools of dcp2-accessible targets for cap-snatching', *Genes and Development*, 27(13), pp. 1511–1525. doi: 10.1101/gad.215384.113.

Houle, D. (1992) 'Comparing evolvability and variability of quantitative traits.', *Genetics*, 130(1), pp. 195–204.

Howick, V. M. and Lazzaro, B. P. (2017) 'The genetic architecture of defence as resistance to and tolerance of bacterial infection in Drosophila melanogaster', *Molecular Ecology*, 26(6), pp. 1533–1546. doi: 10.1111/mec.14017.

Huang, C. *et al.* (2015) 'RNAi pathway participates in chromosome segregation in mammalian cells', *Cell Discovery*. Nature Publishing Group, 1, p. 15029. doi: 10.1038/celldisc.2015.29.

Huang, Z. *et al.* (2013) 'An Antiviral Role for Antimicrobial Peptides during the Arthropod Response to Alphavirus Replication', *Journal of Virology*, 87(8), pp. 4272–4280. doi: 10.1128/JVI.03360-12.

Huber, C. D. *et al.* (2016) 'Detecting recent selective sweeps while controlling for mutation rate and background selection', *Molecular Ecology*, 25(1), pp. 142–156. doi: 10.1111/mec.13351.

Hudson, R. R. (2002) 'Generating samples under a Wright-Fisher neutral model of genetic variation', *Bioinformatics*. Oxford University Press, 18(2), pp. 337–338. doi: 10.1093/bioinformatics/18.2.337.

Huelsenbeck, J. P. and Ronquist, F. (2001) 'MRBAYES: Bayesian inference of phylogenetic trees', *Bioinformatics* , 17(8), pp. 754–755. doi: 10.1093/bioinformatics/17.8.754.

Huger, A. M. (1966) 'A virus disease of the Indian rhinoceros beetle, Oryctes

rhinoceros(Linnaeus), caused by a new type of insect virus, Rhabdionvirus oryctes gen. n., sp. n.', *Journal of invertebrate pathology*, 8(1), pp. 38–51.

Huger, A. M. (1972) 'Grundlagen zur biologischen Bekämpfung des Indischen Nashornkäfers, Oryctes rhinoceros (L.), mit Rhabdionvirus oryctes: Histopathologie der Virose bei Käfern', *Zeitschrift für Angewandte Entomologie*, 72(1–4), pp. 309–319. doi: 10.1111/j.1439-0418.1972.tb02247.x.

Huger, A. M. (1985) 'A new virus disease of crickets (Orthoptera: Gryllidae) causing macronucleosis of fatbody', *Journal of Invertebrate Pathology*. Academic Press, 45(1), pp. 108–111. doi: 10.1016/0022-2011(85)90055-2.

Huger, A. M. (2005) 'The Oryctes virus: its detection, identification, and implementation in biological control of the coconut palm rhinoceros beetle, Oryctes rhinoceros (Coleoptera: Scarabaeidae).', *Journal of invertebrate pathology*, 89(1), pp. 78–84. doi: 10.1016/j.jip.2005.02.010.

Huger, A. M. (2005) 'The Oryctes virus: Its detection, identification, and implementation in biological control of the coconut palm rhinoceros beetle, Oryctes rhinoceros (Coleoptera: Scarabaeidae)', *Journal of Invertebrate Pathology*. Academic Press, 89(1), pp. 78–84. doi: 10.1016/J.JIP.2005.02.010.

Huszar, T. and Imler, J.-L. (2008) 'Drosophila viruses and the study of antiviral host-defense.', *Advances in virus research*, 72, pp. 227–65. doi: 10.1016/S0065-3527(08)00406-5.

Ivanov, D. K. *et al.* (2015) 'Longevity GWAS Using the *Drosophila* Genetic Reference Panel', *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 70(12), pp. 1470–1478. doi: 10.1093/gerona/glv047.

Jayachandran, B., Hussain, M. and Asgari, S. (2012) 'RNA interference as a cellular defense mechanism against the DNA virus baculovirus.', *Journal of virology*, 86(24), pp. 13729–34. doi: 10.1128/JVI.02041-12.

Jeffreys, A. J. *et al.* (1988) 'Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA.', *Nature*, 332(6161), pp. 278–81. doi: 10.1038/332278a0.

Jinek, M. *et al.* (2008) 'The C-terminal region of Ge-1 presents conserved structural features required for P-body localization', *Rna*, 14(10), pp. 1991–1998. doi: 10.1261/rna.1222908.

Jorda, J. and Kajava, A. V (2009) 'T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm.', *Bioinformatics (Oxford, England)*, 25(20), pp. 2632–8. doi: 10.1093/bioinformatics/btp482.

Jouandin, P., Ghiglione, C. and Noselli, S. (2014) 'Starvation induces FoxO-dependent mitotic-to-endocycle switch pausing during Drosophila oogenesis', *Development*, 141(15), pp. 3013–3021. doi: 10.1242/dev.108399.

Juneja, P. and Lazzaro, B. P. (2010) 'Haplotype structure and expression divergence at the Drosophila cellular immune gene eater', *Molecular Biology and Evolution*, 27(10), pp. 2284–2299. doi: 10.1093/molbev/msq114.

Jupatanakul, N. *et al.* (2017) 'Engineered Aedes aegypti JAK/STAT Pathway-Mediated Immunity to Dengue Virus', *PLoS Neglected Tropical Diseases*. Public Library of Science,

11(1), p. e0005187. doi: 10.1371/journal.pntd.0005187.

Kaneko, T. *et al.* (2006) 'PGRP-LC and PGRP-LE have essential yet distinct functions in the drosophila immune response to monomeric DAP-type peptidoglycan', *Nature Immunology*, 7(7), pp. 715–723. doi: 10.1038/ni1356.

Kapun, M. *et al.* (2018) 'Genomic analysis of European Drosophila melanogaster populations on a dense spatial scale reveals longitudinal population structure and continent-wide selection', *bioRxiv*.

Karlas, A. *et al.* (2016) 'A human genome-wide loss-of-function screen identifies effective chikungunya antiviral drugs.', *Nature communications*, 7, p. 11320. doi: 10.1038/ncomms11320.

Kataoka, C. *et al.* (2012) 'Baculovirus GP64-mediated entry into mammalian cells.', *Journal of virology*. American Society for Microbiology, 86(5), pp. 2610–20. doi: 10.1128/JVI.06704-11.

Kawamura, Y. *et al.* (2008) 'Drosophila endogenous small RNAs bind to Argonaute 2 in somatic cells.', *Nature*. Nature Publishing Group, 453(7196), pp. 793–7. doi: 10.1038/nature06938.

Keightley, P. D. *et al.* (2014) 'Estimation of the spontaneous mutation rate per nucleotide site in a Drosophila melanogaster full-sib family.', *Genetics*, 196(1), pp. 313–20. doi: 10.1534/genetics.113.158758.

Keightley, P. D. and Eyre-Walker, A. (2007) 'Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies', *Genetics*, 177(4), pp. 2251–2261. doi: 10.1534/genetics.107.080663.

Keightley, P. D. and Eyre-Walker, A. (2010) 'What can we learn about the distribution of fitness effects of new mutations from DNA sequence data?', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 365(1544), pp. 1187–93. doi: 10.1098/rstb.2009.0266.

Kemp, C. *et al.* (2013) 'Broad RNA Interference-Mediated Antiviral Immunity and Virus-Specific Inducible Responses in Drosophila', *The Journal of Immunology*. NIH Public Access, 190(2), pp. 650–658. doi: 10.4049/jimmunol.1102486.

Khong, A. and Jan, E. (2011) 'Modulation of Stress Granules and P Bodies during Dicistrovirus Infection', *Journal of Virology*. American Society for Microbiology, 85(4), pp. 1439–1451. doi: 10.1128/JVI.02220-10.

Khurana, J. S. *et al.* (2011) 'Adaptation to P element transposon invasion in Drosophila melanogaster.', *Cell*, 147(7), pp. 1551–63. doi: 10.1016/j.cell.2011.11.042.

King, E. G. and Long, A. D. (2017) 'The Beavis Effect in Next-Generation Mapping Panels in Drosophila melanogaster.', *G3 (Bethesda, Md.)*, 7(6), pp. 1643–1652. doi: 10.1534/g3.117.041426.

King, E. G., Macdonald, S. J. and Long, A. D. (2012) 'Properties and power of the Drosophila Synthetic Population Resource for the routine dissection of complex traits.', *Genetics*, 191(3), pp. 935–49. doi: 10.1534/genetics.112.138537.

Klattenhoff, C. and Theurkauf, W. (2008) 'Biogenesis and germline functions of piRNAs.', *Development (Cambridge, England)*, 135(1), pp. 3–9. doi: 10.1242/dev.006486.

Kolaczkowski, B., Hupalo, D. N. and Kern, A. D. (2011) 'Recurrent adaptation in RNA interference genes across the Drosophila phylogeny.', *Molecular biology and evolution*, 28(2), pp. 1033–42. doi: 10.1093/molbev/msq284.

Kousathanas, A., Halligan, D. L. and Keightley, P. D. (2014) 'Faster-X adaptive protein evolution in house mice.', *Genetics*, 196(4), pp. 1131–43. doi: 10.1534/genetics.113.158246.

Kronforst, M. R. *et al.* (2006) 'Multilocus analyses of admixture and introgression among hybridizing Heliconius butterflies', *Evolution*, 60(6), pp. 1254–1268. doi: 10.1111/j.0014-3820.2006.tb01203.x.

Kronforst, M. R. *et al.* (2013) 'Hybridization reveals the evolving genomic architecture of speciation.', *Cell reports*, 5(3), pp. 666–77. doi: 10.1016/j.celrep.2013.09.042.

Kwak, P. B. and Tomari, Y. (2012) 'The N domain of Argonaute drives duplex unwinding during RISC assembly.', *Nature structural & molecular biology*, 19(2), pp. 145–51. doi: 10.1038/nsmb.2232.

Lachaise, D. and Silvain, J.-F. (2004) 'How two Afrotropical endemics made two cosmopolitan human commensals: the Drosophila melanogaster-D. simulans palaeogeographic riddle.', *Genetica*, 120(1–3), pp. 17–39.

Lack, J. B. *et al.* (2016) 'A Thousand Fly Genomes: An Expanded Drosophila Genome Nexus.', *Molecular biology and evolution*. Oxford University Press, 33(12), pp. 3308–3313. doi: 10.1093/molbev/msw195.

Lambrechts, L. *et al.* (2012) 'Specificity of resistance to dengue virus isolates is associated with genotypes of the mosquito antiviral gene Dicer-2', *Proceedings of the Royal Society B: Biological Sciences*, 280(1751), pp. 20122437–20122437. doi: 10.1098/rspb.2012.2437.

Lamiable, O., Arnold, J., *et al.* (2016) 'Analysis of the contribution of hemocytes and autophagy to *Drosophila* antiviral immunity', *Journal of Virology*. Edited by R. M. Sandri-Goldin, 90(March), p. JVI.00238-16. doi: 10.1128/JVI.00238-16.

Lamiable, O., Kellenberger, C., *et al.* (2016) 'Cytokine Diedel and a viral homologue suppress the IMD pathway in *Drosophila*', *Proceedings of the National Academy of Sciences*, 113(3), pp. 698–703. doi: 10.1073/pnas.1516122113.

Lamiable, O. and Imler, J. L. (2014) 'Induced antiviral innate immunity in Drosophila', *Current Opinion in Microbiology*. NIH Public Access, 20, pp. 62–68. doi: 10.1016/j.mib.2014.05.006.

Langmead, B. and Salzberg, S. L. (2012) 'Fast gapped-read alignment with Bowtie 2.', *Nature methods*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 9(4), pp. 357–9. doi: 10.1038/nmeth.1923.

Larracuente, A. M. and Clark, A. G. (2014) 'Recent selection on the Y-to-dot translocation in Drosophila pseudoobscura.', *Molecular biology and evolution*, 31(4), pp. 846–56. doi: 10.1093/molbev/msu002.

Lee, Y. C. G. and Langley, C. H. (2012) 'Long-term and short-term evolutionary impacts of

transposable elements on Drosophila.', *Genetics*. Genetics Society of America, 192(4), pp. 1411–32. doi: 10.1534/genetics.112.145714.

Lee, Y. S. *et al.* (2004) 'Distinct roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA silencing pathways', *Cell*, 117(1), pp. 69–81. doi: 10.1016/S0092-8674(04)00261-2.

Leibfried, A. *et al.* (2008) 'Drosophila Cip4 and WASp define a branch of the Cdc42-Par6-aPKC pathway regulating E-cadherin endocytosis.', *Current biology : CB*, 18(21), pp. 1639–48. doi: 10.1016/j.cub.2008.09.063.

Lentz, T. L. *et al.* (1983) 'The acetylcholine receptor as a cellular receptor for rabies virus.', *The Yale journal of biology and medicine*, 56(4), pp. 315–22.

Lewis, S. H. *et al.* (2016) 'Repeated Duplication of Argonaute2 Is Associated with Strong Selection and Testis Specialization in Drosophila', *Genetics*, 204(2).

Lewis, S. H. *et al.* (2018) 'Pan-arthropod analysis reveals somatic piRNAs as an ancestral defence against transposable elements', *Nature Ecology & Evolution*, 2(1), pp. 174–181. doi: 10.1038/s41559-017-0403-4.

Li, H., Li, W. X. and Ding, S. W. (2002) 'Induction and Suppression of RNA Silencing by an Animal Virus', *Science*, 296(5571), pp. 1319–1321. doi: 10.1126/science.1070948.

Li, Y.-X. and Dijkers, P. F. (2015) 'Specific calcineurin isoforms are involved in Drosophila toll immune signaling.', *Journal of immunology (Baltimore, Md. : 1950)*, 194(1), pp. 168–76. doi: 10.4049/jimmunol.1401080.

Li, Y. *et al.* (2013) 'RNA Interference Functions as an Antiviral Immunity Mechanism in Mammals', *Science*, 342(6155), pp. 231–234. doi: 10.1126/science.1241911.

Liao, Y., Smyth, G. K. and Shi, W. (2013) 'The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote', *Nucleic Acids Research*. ACM, New York, 41(10), pp. e108–e108. doi: 10.1093/nar/gkt214.

Lingel, A. *et al.* (2003) 'Structure and nucleic-acid binding of the Drosophila Argonaute 2 PAZ domain.', *Nature*, 426(6965), pp. 465–9. doi: 10.1038/nature02123.

Liu, B. *et al.* (2013) 'P53-Mediated Rapid Induction of Apoptosis Conveys Resistance to Viral Infection in Drosophila melanogaster', *PLoS Pathogens*. Public Library of Science, 9(2), p. e1003137. doi: 10.1371/journal.ppat.1003137.

Liu, X., Hodgson, J. J. and Buchon, N. (2017) 'Drosophila as a model for homeostatic, antibacterial, and antiviral mechanisms in the gut', *PLOS Pathogens*. Edited by K. A. Kline, 13(5), p. e1006277. doi: 10.1371/journal.ppat.1006277.

Liu, Y. *et al.* (2009) 'C3PO, an Endoribonuclease That Promotes RNAi by Facilitating RISC Activation', *Science*, 325(5941), pp. 750–753. doi: 10.1126/science.1176325.

Loegering, D. J. and Lennartz, M. R. (2011) 'Protein kinase C and toll-like receptor signaling.', *Enzyme research*. Hindawi Publishing Corporation, 2011, p. 537821. doi: 10.4061/2011/537821.

Long, G. *et al.* (2006) 'Functional entry of baculovirus into insect and mammalian cells is dependent on clathrin-mediated endocytosis.', *Journal of virology*. American Society for Microbiology, 80(17), pp. 8830–3. doi: 10.1128/JVI.00880-06.

Longdon, B., Hadfield, J. D., *et al.* (2015) 'The causes and consequences of changes in virulence following pathogen host shifts.', *PLoS pathogens*. Public Library of Science, 11(3), p. e1004728. doi: 10.1371/journal.ppat.1004728.

Longdon, B., Murray, G. G. R., *et al.* (2015) 'The evolution, diversity, and host associations of rhabdoviruses', *Virus Evolution*, 1(1), p. vev014. doi: 10.1093/ve/vev014.

Longdon, B. and Jiggins, F. M. (2012) 'Vertically transmitted viral endosymbionts of insects: do sigma viruses walk alone?', *Proceedings of the Royal Society B: Biological Sciences*. The Royal Society, 279(1744), pp. 3889–3898. doi: 10.1098/rspb.2012.1208.

Longdon, B., Obbard, D. J. and Jiggins, F. M. (2010) 'Sigma viruses from three species of Drosophila form a major new clade in the rhabdovirus phylogeny.', *Proceedings. Biological sciences*, 277(1678), pp. 35–44. doi: 10.1098/rspb.2009.1472.

Love, M. I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.', *Genome biology*, 15(12), p. 550. doi: 10.1186/s13059-014-0550-8.

Lunter, G. and Goodson, M. (2011) 'Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads.', *Genome research*, 21(6), pp. 936–9. doi: 10.1101/gr.111120.110.

Luplertlop, N. *et al.* (2011) 'Induction of a peptide with activity against a broad spectrum of pathogens in the Aedes aegypti salivary gland, following infection with Dengue Virus', *PLoS Pathogens*. Edited by M. S. Diamond, 7(1), p. e1001252. doi: 10.1371/journal.ppat.1001252.

Ma, J.-B. *et al.* (2005) 'Structural basis for 5'-end-specific recognition of guide RNA by the A. fulgidus Piwi protein.', *Nature*, 434(7033), pp. 666–70. doi: 10.1038/nature03514.

Ma, J.-B., Ye, K. and Patel, D. J. (2004) 'Structural basis for overhang-specific small interfering RNA recognition by the PAZ domain.', *Nature*, 429(6989), pp. 318–322. doi: 10.1038/nature02519.

Mackay, T. F. C. *et al.* (2012) 'The Drosophila melanogaster Genetic Reference Panel.', *Nature*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 482(7384), pp. 173–8. doi: 10.1038/nature10811.

Magwire, M. M. *et al.* (2011) 'Successive increases in the resistance of Drosophila to viral infection through a transposon insertion followed by a duplication', *PLoS Genetics*. Edited by D. J. Begun, 7(10), p. e1002337. doi: 10.1371/journal.pgen.1002337.

Magwire, M. M. *et al.* (2012) 'Genome-Wide Association Studies Reveal a Simple Genetic Basis of Resistance to Naturally Coevolving Viruses in Drosophila melanogaster', *PLoS Genetics*, 8(11). doi: 10.1371/journal.pgen.1003057.

Maillard, P. V. *et al.* (2013) 'Antiviral RNA Interference in Mammalian Cells', *Science*, 342(6155), pp. 235–238. doi: 10.1126/science.1241930.

Marçais, G. and Kingsford, C. (2011) 'A fast, lock-free approach for efficient parallel counting of occurrences of k-mers.', *Bioinformatics (Oxford, England)*, 27(6), pp. 764–70. doi: 10.1093/bioinformatics/btr011.

Mariani, R. *et al.* (2003) 'Species-specific exclusion of APOBEC3G from HIV-1 virions by Vif.',

Cell, 114(1), pp. 21–31.

Martin, M. (2011) 'Cutadapt removes adapter sequences from high-throughput sequencing reads', *EMBnet.journal*, 17(1), p. 10. doi: 10.14806/ej.17.1.200.

Martins, N. E. *et al.* (2014) 'Host adaptation to viruses relies on few genes with different cross-resistance properties', *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 111(16), pp. 5938–5943. doi: 10.1073/pnas.1400378111.

McDonald, J. H. and Kreitman, M. (1991) 'Adaptive protein evolution at the Adh locus in Drosophila.', *Nature*, 351(6328), pp. 652–4. doi: 10.1038/351652a0.

McGaugh, S. E. *et al.* (2012) 'Recombination Modulates How Selection Affects Linked Sites in Drosophila', *PLoS Biology*. Edited by N. H. Barton. Public Library of Science, 10(11), p. e1001422. doi: 10.1371/journal.pbio.1001422.

McGaughran, A., Morgan, K. and Sommer, R. J. (2013) 'Unraveling the evolutionary history of the nematode Pristionchus pacificus: from lineage diversification to island colonization.', *Ecology and evolution*, 3(3), pp. 667–75. doi: 10.1002/ece3.495.

McTaggart, S. J. *et al.* (2015) 'Novel insights into the insect trancriptome response to a natural DNA virus', *BMC Genomics*, 16(1), p. 310. doi: 10.1186/s12864-015-1499-z.

Medd, N. C. *et al.* (2018) 'The virome of Drosophila suzukii, an invasive pest of soft fruit.', *Virus evolution*, 4(1), p. vey009. doi: 10.1093/ve/vey009.

Meister, G. (2013) 'Argonaute proteins: functional insights and emerging roles.', *Nature reviews. Genetics*. Nature Publishing Group, 14(7), pp. 447–59. doi: 10.1038/nrg3462.

Melchjorsen, J., Matikainen, S. and Paludan, S. R. (2009) 'Activation and evasion of innate antiviral immunity by herpes simplex virus.', *Viruses*. Multidisciplinary Digital Publishing Institute (MDPI), 1(3), pp. 737–59. doi: 10.3390/v1030737.

Merkling, S. H. *et al.* (2015) 'The epigenetic regulator G9a mediates tolerance to RNA virus infection in Drosophila.', *PLoS pathogens*. Edited by P. F. Vale, 11(4), p. e1004692. doi: 10.1371/journal.ppat.1004692.

Merkling, S. H. and van Rij, R. P. (2013) 'Beyond RNAi: Antiviral defense strategies in Drosophila and mosquito', *Journal of Insect Physiology*, 59(2), pp. 159–170. doi: 10.1016/j.jinsphys.2012.07.004.

van Mierlo, J. T. *et al.* (2012) 'Convergent evolution of argonaute-2 slicer antagonism in two distinct insect RNA viruses.', *PLoS pathogens*. Public Library of Science, 8(8), p. e1002872. doi: 10.1371/journal.ppat.1002872.

van Mierlo, J. T. *et al.* (2012) 'Convergent Evolution of Argonaute-2 Slicer Antagonism in Two Distinct Insect RNA Viruses', *PLoS Pathogens*, 8(8). doi: 10.1371/journal.ppat.1002872.

van Mierlo, J. T. *et al.* (2014) 'Novel Drosophila Viruses Encode Host-Specific Suppressors of RNAi', *PLoS Pathogens*, 10(7). doi: 10.1371/journal.ppat.1004256.

Miesen, P. *et al.* (2016) 'Small RNA Profiling in Dengue Virus 2-Infected Aedes Mosquito Cells Reveals Viral piRNAs and Novel Host miRNAs.', *PLoS neglected tropical diseases*, 10(2), p. e0004452. doi: 10.1371/journal.pntd.0004452.

Miesen, P., Girardi, E. and van Rij, R. P. (2015) 'Distinct sets of PIWI proteins produce

arbovirus and transposon-derived piRNAs in Aedes aegypti mosquito cells.', *Nucleic acids research*, 43(13), pp. 6545–56. doi: 10.1093/nar/gkv590.

Mohn, F., Handler, D. and Brennecke, J. (2015) 'piRNA-guided slicing specifies transcripts for Zucchini-dependent, phased piRNA biogenesis.', *Science (New York, N.Y.)*, 348(6236), pp. 812–7. doi: 10.1126/science.aaa1039.

Molleston, J. and Cherry, S. (2017) 'Attacked from All Sides: RNA Decay in Antiviral Defense', *Viruses*. Multidisciplinary Digital Publishing Institute, 9(1), p. 2. doi: 10.3390/v9010002.

Molleston, J. M. *et al.* (2016) 'A conserved virus-induced cytoplasmic TRAMP-like complex recruits the exosome to target viral RNA for degradation', *Genes and Development*, 30(14), pp. 1658–1670. doi: 10.1101/gad.284604.116.

Moon, S. L. *et al.* (2012) 'A noncoding RNA produced by arthropod-borne flaviviruses inhibits the cellular exoribonuclease XRN1 and alters host mRNA stability', *Rna*. Cold Spring Harbor Laboratory Press, 18(11), pp. 2029–2040. doi: 10.1261/rna.034330.112.

Moon, S. L. and Wilusz, J. (2013) 'Cytoplasmic Viruses: Rage against the (Cellular RNA Decay) Machine', *PLoS Pathogens*. Edited by V. Racaniello. Public Library of Science, 9(12), pp. 1–3. doi: 10.1371/journal.ppat.1003762.

Morazzani, E. M. *et al.* (2012) 'Production of Virus-Derived Ping-Pong-Dependent piRNA-like Small RNAs in the Mosquito Soma', *PLoS Pathogens*. Edited by S.-W. Ding. Public Library of Science, 8(1), p. e1002470. doi: 10.1371/journal.ppat.1002470.

Moy, R. H. *et al.* (2014) 'Stem-loop recognition by DDX17 facilitates miRNA processing and antiviral defense', *Cell*, 158(4), pp. 764–777. doi: 10.1016/j.cell.2014.06.023.

Moy, R. H. and Cherry, S. (2013) 'Antimicrobial autophagy: A conserved innate immune response in drosophila', *Journal of Innate Immunity*. NIH Public Access, 5(5), pp. 444–455. doi: 10.1159/000350326.

Muerdter, F. *et al.* (2013) 'A genome-wide RNAi screen draws a genetic framework for transposon control and primary piRNA biogenesis in Drosophila.', *Molecular cell*, 50(5), pp. 736–48. doi: 10.1016/j.molcel.2013.04.006.

Myllymaki, H., Valanne, S. and Ramet, M. (2014) 'The Drosophila Imd Signaling Pathway', *The Journal of Immunology*, 192(8), pp. 3455–3462. doi: 10.4049/jimmunol.1303309.

Nainu, F. *et al.* (2015) 'Protection of Insects against Viral Infection by Apoptosis-Dependent Phagocytosis', *The Journal of Immunology*, 195(12), pp. 5696–5706. doi: 10.4049/jimmunol.1500613.

Nakamoto, M. *et al.* (2012) 'Virus Recognition by Toll-7 Activates Antiviral Autophagy in Drosophila', *Immunity*. NIH Public Access, 36(4), pp. 658–667. doi: 10.1016/j.immuni.2012.03.003.

Nan, Y., Wu, C. and Zhang, Y.-J. (2017) 'Interplay between Janus Kinase/Signal Transducer and Activator of Transcription Signaling Activated by Type I Interferons and Viral Antagonism.', *Frontiers in immunology*, 8, p. 1758. doi: 10.3389/fimmu.2017.01758.

Nayak, A. *et al.* (2010) 'Cricket paralysis virus antagonizes Argonaute 2 to modulate antiviral defense in Drosophila', *Nature Structural & Molecular Biology*, 17(5), pp. 547–554. doi:

10.1038/nsmb.1810.

Nguyen, Q., Nielsen, L. K. and Reid, S. (2013) 'Genome scale transcriptomics of baculovirus-insect interactions.', *Viruses*. Multidisciplinary Digital Publishing Institute (MDPI), 5(11), pp. 2721–47. doi: 10.3390/v5112721.

Nielsen, R. (2005) 'Genomic scans for selective sweeps using SNP data', *Genome Research*, 15(11), pp. 1566–1575. doi: 10.1101/gr.4252305.

Nielsen, R. *et al.* (2005) 'Genomic scans for selective sweeps using SNP data', *Genome Research*, 15(11), pp. 1566–1575. doi: 10.1101/gr.4252305.

Nitta, K. R. *et al.* (2015) 'Conservation of transcription factor binding specificities across 600 million years of bilateria evolution.', *eLife*, 4. doi: 10.7554/eLife.04837.

Noland, J. *et al.* (2013) 'Gut Transcription in Helicoverpa zea is Dynamically Altered in Response to Baculovirus Infection', *Insects*. Multidisciplinary Digital Publishing Institute, 4(3), pp. 506–520. doi: 10.3390/insects4030506.

Nusslein-Volhard, C. (1979) 'Maternal effect mutations that affect the spatial coordinates of the embryo of Drosophila melanogaster.', *Determinants of spatial organization.*, pp. 185–211.

Obbard, D. J. *et al.* (2006) 'Natural selection drives extremely rapid evolution in antiviral RNAi genes', *Current Biology*, 16(6), pp. 580–585. doi: 10.1016/j.cub.2006.01.065.

Obbard, D. J. *et al.* (2007) 'Population genetics of Plasmodium resistance genes in Anopheles gambiae: no evidence for strong selection.', *Molecular ecology*, 16(16), pp. 3497–510. doi: 10.1111/j.1365-294X.2007.03395.x.

Obbard, D. J. *et al.* (2009) 'Quantifying adaptive evolution in the Drosophila immune system', *PLoS Genetics*, 5(10). doi: 10.1371/journal.pgen.1000698.

Obbard, D. J. *et al.* (2009) 'The evolution of RNAi as a defence against viruses and transposable elements.', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 364(1513), pp. 99–115. doi: 10.1098/rstb.2008.0168.

Obbard, D. J. *et al.* (2011) 'Recent and recurrent selective sweeps of the antiviral RNAi gene argonaute-2 in three species of drosophila', *Molecular Biology and Evolution*, 28(2), pp. 1043–1056. doi: 10.1093/molbev/msq280.

Obbard, D. J. *et al.* (2012) 'Estimating Divergence Dates and Substitution Rates in the Drosophila Phylogeny', *Molecular Biology and Evolution*. Oxford University Press, 29(11), pp. 3459–3473. doi: 10.1093/molbev/mss150.

Obbard, D. J. and Dudas, G. (2014) 'The genetics of host-virus coevolution in invertebrates', *Current Opinion in Virology*, 8, pp. 73–78. doi: 10.1016/j.coviro.2014.07.002.

Okamura, K. *et al.* (2004) 'Distinct roles for Argonaute proteins in small RNA-directed RNA cleavage pathways.', *Genes & development*. Cold Spring Harbor Laboratory Press, 18(14), pp. 1655–66. doi: 10.1101/gad.1210204.

Orban, T. I. and Izaurralde, E. (2005) 'Decay of mRNAs targeted by RISC requires XRN1, the Ski complex, and the exosome.', *RNA (New York, N.Y.)*, 11(4), pp. 459–69. doi: 10.1261/rna.7231505.

Palmer, W. H. *et al.* (2018) 'Isolation of a natural DNA virus of Drosophila melanogaster, and characterisation of host resistance and immune responses', *PLOS Pathogens*. Edited by J.-L. Imler, 14(6), p. e1007050. doi: 10.1371/journal.ppat.1007050.

Palmer, W. H., Hadfield, J. D. and Obbard, D. J. (2018) 'RNA Interference Pathways Display High Rates of Adaptive Protein Evolution in Multiple Invertebrates', *Genetics*. Genetics, p. genetics.300567.2017. doi: 10.1534/genetics.117.300567.

Palmer, W. H. and Obbard, D. J. (2016) 'Variation and Evolution in the Glutamine-Rich Repeat Region of Drosophila Argonaute-2', *G3: Genes, Genomes, Genetics*, 6(8), pp. 2563–2572. doi: 10.1534/g3.116.031880.

Palmer, W. H., Varghese, F. S. and van Rij, R. P. (2018) 'Natural Variation in Resistance to Virus Infection in Dipteran Insects.', *Viruses*, 10(3), p. 118. doi: 10.3390/v10030118.

Paradkar, P. N. *et al.* (2012) 'Secreted Vago restricts West Nile virus infection in Culex mosquito cells by activating the Jak-STAT pathway', *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 109(46), pp. 18915–18920. doi: 10.1073/pnas.1205231109.

Pardo-Diaz, C. *et al.* (2012) 'Adaptive introgression across species boundaries in Heliconius butterflies.', *PLoS genetics*. Public Library of Science, 8(6), p. e1002752. doi: 10.1371/journal.pgen.1002752.

Parisien, J.-P., Lau, J. F. and Horvath, C. M. (2002) 'STAT2 acts as a host range determinant for species-specific paramyxovirus interferon antagonism and simian virus 5 replication.', *Journal of virology*, 76(13), pp. 6435–41.

Parker, J. S., Roe, S. M. and Barford, D. (2004) 'Crystal structure of a PIWI protein suggests mechanisms for siRNA recognition and slicer activity.', *The EMBO journal*. EMBO Press, 23(24), pp. 4727–37. doi: 10.1038/sj.emboj.7600488.

Paterson, S. *et al.* (2010) 'Antagonistic coevolution accelerates molecular evolution', *Nature*. Nature Publishing Group, 464(7286), pp. 275–278. doi: 10.1038/nature08798.

Pavlidis, P. *et al.* (2013) 'SweeD: Likelihood-Based Detection of Selective Sweeps in Thousands of Genomes', *Molecular Biology and Evolution*, 30(9), pp. 2224–2234. doi: 10.1093/molbev/mst112.

Payne, C. C. (1974) 'The Isolation and Characterization of a Virus from Oryctes rhinoceros', *Journal of General Virology*, 25(1), pp. 105–116. doi: 10.1099/0022-1317-25-1-105.

Peccoud, J. *et al.* (2017) 'Massive horizontal transfer of transposable elements in insects.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 114(18), pp. 4721–4726. doi: 10.1073/pnas.1621178114.

Petrov, D. A. *et al.* (1995) 'Diverse transposable elements are mobilized in hybrid dysgenesis in Drosophila virilis.', *Proceedings of the National Academy of Sciences*, 92(17), pp. 8050–8054. doi: 10.1073/pnas.92.17.8050.

Pfeiffenberger, C. *et al.* (2010) 'Locomotor activity level monitoring using the Drosophila Activity Monitoring (DAM) System.', *Cold Spring Harbor protocols*, 2010(11), p. pdb.prot5518.

Pham, L. N. *et al.* (2007) 'A Specific Primed Immune Response in Drosophila Is Dependent

on Phagocytes', *PLoS Pathogens*. Public Library of Science, 3(3), p. e26. doi: 10.1371/journal.ppat.0030026.

Piontkivska, H. *et al.* (2016) 'Role of host-driven mutagenesis in determining genome evolution of sigma virus (DMelSV; Rhabdoviridae) in drosophila melanogaster', *Genome Biology and Evolution*. Oxford University Press, 8(9), pp. 2952–2963. doi: 10.1093/gbe/evw212.

Poirier, E. Z. *et al.* (2018) 'Dicer-2-Dependent Generation of Viral DNA from Defective Genomes of RNA Viruses Modulates Antiviral Immunity in Insects.', *Cell host & microbe*, 23(3), p. 353–365.e8. doi: 10.1016/j.chom.2018.02.001.

Pombi, M. *et al.* (2006) 'Variation in recombination rate across the X chromosome of Anopheles gambiae.', *The American journal of tropical medicine and hygiene*, 75(5), pp. 901–3.

Qiu, P., Pan, P. C. and Govind, S. (1998) 'A role for the Drosophila Toll/Cactus pathway in larval hematopoiesis.', *Development (Cambridge, England)*, 125(10), pp. 1909–20.

Raina, A. K. and Adams, J. R. (1995) 'Gonad-specific virus of corn earworm', *Nature*. Nature Publishing Group, 374(6525), pp. 770–770. doi: 10.1038/374770a0.

Rajsbaum, R. *et al.* (2012) 'Species-specific inhibition of RIG-I ubiquitination and IFN induction by the influenza A virus NS1 protein.', *PLoS pathogens*. Edited by A. Pekosz, 8(11), p. e1003059. doi: 10.1371/journal.ppat.1003059.

Ralston, A. L., Huang, Y. S. and Kawanishi, C. Y. (1981) 'Cell culture studies with the IMC-Hz-1 nonoccluded virus.', *Virology*, 115(1), pp. 33–44.

Ramirez, J. L. and Dimopoulos, G. (2010) 'The Toll immune signaling pathway control conserved anti-dengue defenses across diverse Ae. aegypti strains and against multiple dengue virus serotypes', *Developmental & Comparative Immunology*, 34(6), pp. 625–629. doi: 10.1016/j.dci.2010.01.006.

Rice, C. *et al.* (2015) 'The Nature, Extent, and Consequences of Genetic Variation in the opa Repeats of Notch in Drosophila.', *G3 (Bethesda, Md.)*, 5(11), pp. 2405–19. doi: 10.1534/g3.115.021659.

Van Rij, R. P. *et al.* (2006) 'The RNA silencing endonuclease Argonaute 2 mediates specific antiviral immunity in Drosophila melanogaster', *Genes and Development*, 20(21), pp. 2985–2995. doi: 10.1101/gad.1482006.

Rödelsperger, C. *et al.* (2014) 'Characterization of genetic diversity in the nematode Pristionchus pacificus from population-scale resequencing data.', *Genetics*, 196(4), pp. 1153–65. doi: 10.1534/genetics.113.159855.

Ross, J. A. *et al.* (2011) 'Caenorhabditis briggsae recombinant inbred line genotypes reveal inter-strain incompatibility and the evolution of recombination.', *PLoS genetics*, 7(7), p. e1002174. doi: 10.1371/journal.pgen.1002174.

Rozhkov, N. V *et al.* (2010) 'Small RNA-based silencing strategies for transposons in the process of invading Drosophila species.', *RNA (New York, N.Y.)*, 16(8), pp. 1634–45. doi: 10.1261/rna.2217810.

Rutschmann, S. *et al.* (2000) 'The Rel protein DIF mediates the antifungal but not the

antibacterial host defense in Drosophila.', *Immunity*, 12(5), pp. 569–80.

Sabeti, P. C. *et al.* (2002) 'Detecting recent positive selection in the human genome from haplotype structure.', *Nature*, 419(6909), pp. 832–7. doi: 10.1038/nature01140.

Sackton, T. B. *et al.* (2007) 'Dynamic evolution of the innate immune system in Drosophila', *Nature Genetics*, 39(12), pp. 1461–1468. doi: 10.1038/ng.2007.60.

Sackton, T. B. *et al.* (2014) 'Positive selection drives faster-Z evolution in silkmoths.', *Evolution; international journal of organic evolution*, 68(8), pp. 2331–42. doi: 10.1111/evo.12449.

Saleh, M.-C. *et al.* (2006) 'The endocytic pathway mediates cell entry of dsRNA to induce RNAi silencing.', *Nature cell biology*. NIH Public Access, 8(8), pp. 793–802. doi: 10.1038/ncb1439.

Saleh, M.-C. *et al.* (2009) 'Antiviral immunity in Drosophila requires systemic RNA interference spread.', *Nature*. NIH Public Access, 458(7236), pp. 346–50. doi: 10.1038/nature07712.

Salichs, E. *et al.* (2009) 'Genome-wide analysis of histidine repeats reveals their role in the localization of human proteins to the nuclear speckles compartment.', *PLoS genetics*. Public Library of Science, 5(3), p. e1000397. doi: 10.1371/journal.pgen.1000397.

Samuel, C. E. (2012) 'ADARs: Viruses and innate immunity', *Current Topics in Microbiology and Immunology*. NIH Public Access, 353(1), pp. 163–195. doi: 10.1007/82_2011_148.

Sansone, C. L. *et al.* (2015) 'Microbiota-dependent priming of antiviral intestinal immunity in Drosophila', *Cell Host and Microbe*. NIH Public Access, 18(5), pp. 571–581. doi: 10.1016/j.chom.2015.10.010.

Sarkies, P. *et al.* (2015) 'Ancient and Novel Small RNA Pathways Compensate for the Loss of piRNAs in Multiple Independent Nematode Lineages', *PLOS Biology*. Edited by L. D. Hurst. Public Library of Science, 13(2), p. e1002061. doi: 10.1371/journal.pbio.1002061.

Sasaki, A. (2000) 'Host-parasite coevolution in a multilocus gene-for-gene system.', *Proceedings. Biological sciences / The Royal Society*, 267(1458), pp. 2183–8. doi: 10.1098/rspb.2000.1267.

Sawaya, S., Jones, M. and Keller, M. (2016) 'Linkage disequilibrium between single nucleotide polymorphisms and hypermutable loci', *Genetics*.

Sawyer, S. A. *et al.* (2003) 'Bayesian Analysis Suggests that Most Amino Acid Replacements in Drosophila Are Driven by Positive Selection', *Journal of Molecular Evolution*. Springer-Verlag, 57(0), pp. S154–S164. doi: 10.1007/s00239-003-0022-3.

Sawyer, S. L. and Elde, N. C. (2012) 'A cross-species view on viruses.', *Current opinion in virology*, 2(5), pp. 561–8. doi: 10.1016/j.coviro.2012.07.003.

Schlenke, T. A. and Begun, D. J. (2003) 'Natural selection drives Drosophila immune system evolution', *Genetics*, 164(4), pp. 1471–1480.

Schoenberg, D. R. and Maquat, L. E. (2012) 'Regulation of cytoplasmic mRNA decay', *Nature Reviews Genetics*, 13(4), pp. 246–259. doi: 10.1038/nrg3160.

Schrider, D. R. *et al.* (2015) 'Soft shoulders ahead: spurious signatures of soft and partial

selective sweeps result from linked hard sweeps.', *Genetics*, 200(1), pp. 267–84. doi: 10.1534/genetics.115.174912.

Sessions, O. M. *et al.* (2009) 'Discovery of insect and human dengue virus host factors', *Nature*. NIH Public Access, 458(7241), pp. 1047–1050. doi: 10.1038/nature07967.

Settles, E. W. and Friesen, P. D. (2008) 'Flock House Virus Induces Apoptosis by Depletion of Drosophila Inhibitor-of-Apoptosis Protein DIAP1', *Journal of Virology*, 82(3), pp. 1378–1388. doi: 10.1128/JVI.01941-07.

Sharma, S. *et al.* (2003) 'Triggering the Interferon Antiviral Response Through an IKK-Related Pathway', *Science*, 300(5622), pp. 1148–1151. doi: 10.1126/science.1081315.

Shelly, S. *et al.* (2009) 'Autophagy is an essential component of Drosophila immunity against vesicular stomatitis virus.', *Immunity*. NIH Public Access, 30(4), pp. 588–98. doi: 10.1016/j.immuni.2009.02.009.

Shoemaker, C. J. and Green, R. (2012) 'Translation drives mRNA quality control', *Nature Structural and Molecular Biology*. NIH Public Access, 19(6), pp. 594–601. doi: 10.1038/nsmb.2301.

Silverman, N. *et al.* (2003) 'Immune activation of NF-kappaB and JNK requires Drosophila TAK1.', *The Journal of biological chemistry*, 278(49), pp. 48928–34. doi: 10.1074/jbc.M304802200.

Sims, G. E. *et al.* (2009) 'Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions', *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 106(8), pp. 2677–2682. doi: 10.1073/pnas.0813249106.

Slater, G. S. C. and Birney, E. (2005) 'Automated generation of heuristics for biological sequence comparison.', *BMC bioinformatics*, 6(1), p. 31. doi: 10.1186/1471-2105-6-31.

Smith, J. M. and Haigh, J. (1974) 'The hitch-hiking effect of a favourable gene.', *Genetical research*, 23(1), pp. 23–35.

Smith, N. G. C. and Eyre-Walker, A. (2002) 'Adaptive protein evolution in Drosophila.', *Nature*, 415(6875), pp. 1022–4. doi: 10.1038/4151022a.

Smits, A. H. *et al.* (2013) 'Stoichiometry of chromatin-associated protein complexes revealed by label-free quantitative mass spectrometry-based proteomics', *Nucleic Acids Research*, 41(1), pp. e28–e28. doi: 10.1093/nar/gks941.

Song, J.-J. *et al.* (2004) 'Crystal structure of Argonaute and its implications for RISC slicer activity.', *Science (New York, N.Y.)*, 305(5689), pp. 1434–7. doi: 10.1126/science.1102514.

Souza-Neto, J. A., Sim, S. and Dimopoulos, G. (2009) 'An evolutionary conserved function of the JAK-STAT pathway in anti-dengue defense', *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 106(42), pp. 17841–17846. doi: 10.1073/pnas.0905006106.

Stabell, A. C. *et al.* (2018) 'Dengue viruses cleave STING in humans but not in nonhuman primates, their presumed natural reservoir.', *eLife*. eLife Sciences Publications, Ltd, 7. doi: 10.7554/eLife.31919.

Steller, H. (2008) 'Regulation of apoptosis in Drosophila', *Cell Death and Differentiation*. Nature Publishing Group, 15(7), pp. 1132–1138. doi: 10.1038/cdd.2008.50.

Stenseth, N. C. and Smith, J. M. (1984) 'Coevolution in Ecosystems: Red Queen Evolution or Stasis?', *Evolution*, 38(4), p. 870. doi: 10.2307/2408397.

Stephan, W. and Li, H. (2007) 'The recent demographic and adaptive history of Drosophila melanogaster.', *Heredity*. Nature Publishing Group, 98(2), pp. 65–8. doi: 10.1038/sj.hdy.6800901.

Stump, A. D. *et al.* (2007) 'Genetic exchange in 2La inversion heterokaryotypes of Anopheles gambiae.', *Insect molecular biology*, 16(6), pp. 703–9. doi: 10.1111/j.1365-2583.2007.00764.x.

Swarts, D. C. *et al.* (2014) 'The evolutionary journey of Argonaute proteins.', *Nature structural & molecular biology*, 21(9), pp. 743–53. doi: 10.1038/nsmb.2879.

Tassetto, M., Kunitomi, M. and Andino, R. (2017) 'Circulating Immune Cells Mediate a Systemic RNAi-Based Adaptive Antiviral Response in Drosophila', *Cell*, 169(2), p. 314–325.e13. doi: 10.1016/j.cell.2017.03.033.

Tauszig, S. *et al.* (2000) 'Toll-related receptors and the control of antimicrobial peptide expression in Drosophila.', *Proceedings of the National Academy of Sciences of the United States of America*, 97(19), pp. 10520–5. doi: 10.1073/pnas.180130797.

Teixeira, L., Ferreira, A. and Ashburner, M. (2008) 'The bacterial symbiont Wolbachia induces resistance to RNA viral infections in Drosophila melanogaster.', *PLoS biology*. Edited by L. Keller, 6(12), p. e2. doi: 10.1371/journal.pbio.1000002.

Thézé, J. *et al.* (2011) 'Paleozoic origin of insect large dsDNA viruses.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 108(38), pp. 15931–5. doi: 10.1073/pnas.1105580108.

Thoetkiattikul, H., Beck, M. H. and Strand, M. R. (2005) 'Inhibitor B-like proteins from a polydnavirus inhibit NF- B activation and suppress the insect immune response', *Proceedings of the National Academy of Sciences*, 102(32), pp. 11426–11431. doi: 10.1073/pnas.0505240102.

Thomas, C. G. *et al.* (2015) 'Full-genome evolutionary histories of selfing, splitting, and selection in Caenorhabditis.', *Genome research*, 25(5), pp. 667–78. doi: 10.1101/gr.187237.114.

Thomson, T. and Lin, H. (2009) 'The biogenesis and function of PIWI proteins and piRNAs: progress and prospect.', *Annual review of cell and developmental biology*, 25, pp. 355–76. doi: 10.1146/annurev.cellbio.24.110707.175327.

Treangen, T. J. and Salzberg, S. L. (2012) 'Repetitive DNA and next-generation sequencing: computational challenges and solutions.', *Nature reviews. Genetics*, 13(1), pp. 36–46. doi: 10.1038/nrg3117.

Unckless, R. L. (2011) 'A DNA Virus of Drosophila', *PLoS ONE*. Edited by R. DeSalle. Public Library of Science, 6(10), p. e26564. doi: 10.1371/journal.pone.0026564.

Unckless, R. L., Howick, V. M. and Lazzaro, B. P. (2016) 'Convergent Balancing Selection on an Antimicrobial Peptide in Drosophila', *Current Biology*, 26(2), pp. 257–262. doi: 10.1016/j.cub.2015.11.063.

Unckless, R. L. and Lazzaro, B. P. (2016) 'The potential for adaptive maintenance of diversity

in insect antimicrobial peptides.', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 371(1695), p. 20150291-. doi: 10.1098/rstb.2015.0291.

Valanne, S., Wang, J.-H. and Ramet, M. (2011) 'The Drosophila Toll Signaling Pathway', *The Journal of Immunology*, 186(2), pp. 649–656. doi: 10.4049/jimmunol.1002302.

Vandergaast, R. *et al.* (2011) 'Active Depletion of Host Cell Inhibitor-of-Apoptosis Proteins Triggers Apoptosis upon Baculovirus DNA Replication', *Journal of Virology*. American Society for Microbiology (ASM), 85(16), pp. 8348–8358. doi: 10.1128/JVI.00667-11.

Vandergaast, R. *et al.* (2015) 'Insect Inhibitor-of-Apoptosis (IAP) Proteins Are Negatively Regulated by Signal-Induced N-Terminal Degrons Absent within Viral IAP Proteins', *Journal of Virology*. American Society for Microbiology (ASM), 89(8), pp. 4481–4493. doi: 10.1128/JVI.03659-14.

Varjak, M. *et al.* (2013) 'Magnetic fractionation and proteomic dissection of cellular organelles occupied by the late replication complexes of Semliki Forest virus.', *Journal of virology*. American Society for Microbiology (ASM), 87(18), pp. 10295–312. doi: 10.1128/JVI.01105-13.

Vermaak, D., Henikoff, S. and Malik, H. S. (2005) 'Positive selection drives the evolution of rhino, a member of the heterochromatin protein 1 family in Drosophila.', *PLoS genetics*. Public Library of Science, 1(1), pp. 96–108. doi: 10.1371/journal.pgen.0010009.

Viswanatha, R. *et al.* (2018) 'Pooled genome-wide CRISPR screening for basal and context-specific fitness gene essentiality in Drosophila cells', *bioRxiv*. Cold Spring Harbor Laboratory, p. 274464. doi: 10.1101/274464.

Vogler, H. *et al.* (2007) 'Modification of small RNAs associated with suppression of RNA silencing by tobamovirus replicase protein.', *Journal of virology*, 81(19), pp. 10379–88. doi: 10.1128/JVI.00727-07.

Voight, B. F. *et al.* (2006) 'A map of recent positive selection in the human genome.', *PLoS biology*. Public Library of Science, 4(3), p. e72. doi: 10.1371/journal.pbio.0040072.

Waldock, J., Olson, K. E. and Christophides, G. K. (2012) 'Anopheles gambiae antiviral immune response to systemic O'nyong-nyong infection.', *PLoS neglected tropical diseases*. Public Library of Science, 6(3), p. e1565. doi: 10.1371/journal.pntd.0001565.

Wallberg, A. *et al.* (2014) 'A worldwide survey of genome sequence variation provides insight into the evolutionary history of the honeybee Apis mellifera.', *Nature genetics*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 46(10), pp. 1081–8. doi: 10.1038/ng.3077.

Wang, J. B., Lu, H.-L. and St. Leger, R. J. (2017) 'The genetic basis for variation in resistance to infection in the Drosophila melanogaster genetic reference panel', *PLOS Pathogens*. Edited by A. Andrianopoulos. Public Library of Science, 13(3), p. e1006260. doi: 10.1371/journal.ppat.1006260.

Wang, X.-H. *et al.* (2006) 'RNA interference directs innate immunity against viruses in adult Drosophila.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 312(5772), pp. 452–4. doi: 10.1126/science.1125694.

Wang, Y. and Jehle, J. A. (2009) 'Nudiviruses and other large, double-stranded circular DNA viruses of invertebrates: New insights on an old topic', *Journal of Invertebrate Pathology*,

101(3), pp. 187–193. doi: 10.1016/j.jip.2009.03.013.

Wayne, M. L., Contamine, D. and Kreitman, M. (1996) 'Molecular population genetics of ref(2)P, a locus which confers viral resistance in Drosophila.', *Molecular biology and evolution*, 13(1), pp. 191–9. doi: 10.1093/oxfordjournals.molbev.a025555.

Webster, C. L. *et al.* (2015) 'The Discovery, Distribution, and Evolution of Viruses Associated with Drosophila melanogaster.', *PLoS biology*. Public Library of Science, 13(7), p. e1002210. doi: 10.1371/journal.pbio.1002210.

Webster, C. L. *et al.* (2016) *Twenty five new viruses associated with the Drosophilidae (Diptera)*, *bioRxiv*. Cold Spring Harbor Labs Journals. doi: 10.1101/041665.

Welch, J. J. (2006) 'Estimating the genomewide rate of adaptive protein evolution in Drosophila.', *Genetics*, 173(2), pp. 821–37. doi: 10.1534/genetics.106.056911.

Weller, A. M. *et al.* (2014) 'Opposing forces of A/T-biased mutations and G/C-biased gene conversions shape the genome of the nematode Pristionchus pacificus.', *Genetics*, 196(4), pp. 1145–52. doi: 10.1534/genetics.113.159863.

West, C. and Silverman, N. (2018) 'p38b and JAK-STAT signaling protect against Invertebrate iridescent virus 6 infection in Drosophila.', *PLoS pathogens*. Edited by S. Cherry, 14(5), p. e1007020. doi: 10.1371/journal.ppat.1007020.

Whitfield, C. W. *et al.* (2006) 'Thrice out of Africa: ancient and recent expansions of the honey bee, Apis mellifera.', *Science (New York, N.Y.)*, 314(5799), pp. 642–5. doi: 10.1126/science.1132772.

Wilfert, L. and Jiggins, F. M. (2010) 'Disease association mapping in Drosophila can be replicated in the wild.', *Biology letters*. The Royal Society, 6(5), pp. 666–8. doi: 10.1098/rsbl.2010.0329.

Wilfert, L. and Jiggins, F. M. (2013) 'The dynamics of reciprocal selective sweeps of host resistance and a parasite counter-adaptation in drosophila', *Evolution*, 67(3), pp. 761–773. doi: 10.1111/j.1558-5646.2012.01832.x.

Williams, T. (2008) 'Natural invertebrate hosts of iridoviruses (Iridoviridae)', *Neotropical Entomology*. Sociedade Entomológica do Brasil, 37(6), pp. 615–632. doi: 10.1590/S1519-566X2008000600001.

Wu, Y.-L., Wu, C. P., Liu, C. Y. Y., Hsu, P. W.-C., *et al.* (2011) 'A non-coding RNA of insect HzNV-1 virus establishes latent viral infection through microRNA.', *Scientific reports*. Nature Publishing Group, 1, p. 60. doi: 10.1038/srep00060.

Wu, Y.-L., Wu, C. P., Liu, C. Y. Y., Lee, S.-T., *et al.* (2011) 'Heliothis zea nudivirus 1 gene hhi1 induces apoptosis which is blocked by the Hz-iap2 gene and a noncoding gene, pag1.', *Journal of virology*. American Society for Microbiology, 85(14), pp. 6856–66. doi: 10.1128/JVI.01843-10.

Xi, Z., Ramirez, J. L. and Dimopoulos, G. (2008) 'The Aedes aegypti toll pathway controls dengue virus infection', *PLoS Pathogens*. Edited by D. S. Schneider. Public Library of Science, 4(7), p. e1000098. doi: 10.1371/journal.ppat.1000098.

Xia Qingyou, et al (2009) 'Complete Resequencing of 40 Genomes Reveals Domestication Events and Genes in Silkworm (Bombyx)', *Science*, 326(5951), pp. 433–436. doi:

10.1126/science.1175151.

Xu, J. *et al.* (2013) 'ERK signaling couples nutrient status to antiviral defense in the insect gut', *Proceedings of the National Academy of Sciences*, 110(37), pp. 15025–15030. doi: 10.1073/pnas.1303193110.

Xu, J. and Cherry, S. (2014) 'Viruses and antiviral immunity in Drosophila', *Developmental and Comparative Immunology*, 42(1), pp. 67–84. doi: 10.1016/j.dci.2013.05.002.

Xue, D. and Robert Horvitz, H. (1995) 'Inhibition of the Caenorhabditis elegans cell-death protease CED-3 by a CED-3 cleavage site in baculovirus p35 protein', *Nature*, 377(6546), pp. 248–251. doi: 10.1038/377248a0.

Yamamoto, K. *et al.* (2008) 'A BAC-based integrated linkage map of the silkworm Bombyx mori.', *Genome biology*. BioMed Central Ltd, 9(1), p. R21. doi: 10.1186/gb-2008-9-1-r21.

Yang, S.-Y. *et al.* (2014) 'Demographic history and gene flow during silkworm domestication.', *BMC evolutionary biology*. BioMed Central Ltd, 14(1), p. 185. doi: 10.1186/s12862-014-0185-0.

Yang, Z. (2007) 'PAML 4: Phylogenetic Analysis by Maximum Likelihood', *Molecular Biology and Evolution*. Oxford University Press, 24(8), pp. 1586–1591. doi: 10.1093/molbev/msm088.

Yang, Z. and Nielsen, R. (2000) 'Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models.', *Molecular biology and evolution*, 17(1), pp. 32–43.

Yasunaga, A. *et al.* (2014) 'Genome-Wide RNAi Screen Identifies Broadly-Acting Host Factors That Inhibit Arbovirus Infection', *PLoS Pathogens*. Edited by A. L. Brass. Public Library of Science, 10(2), p. e1003914. doi: 10.1371/journal.ppat.1003914.

Ye, T. and Zhang, X. (2013) 'Involvement of Ran in the regulation of phagocytosis against virus infection in S2 cells', *Developmental and Comparative Immunology*, 41(4), pp. 491–497. doi: 10.1016/j.dci.2013.07.015.

Yi, M. *et al.* (2014) 'Rapid evolution of piRNA pathway in the teleost fish: implication for an adaptation to transposon diversity.', *Genome biology and evolution*. Oxford University Press, 6(6), pp. 1393–407. doi: 10.1093/gbe/evu105.

Young, M. D. *et al.* (2010) 'Gene ontology analysis for RNA-seq: accounting for selection bias.', *Genome biology*, 11(2), p. R14. doi: 10.1186/gb-2010-11-2-r14.

Yu, J. H. *et al.* (2005) 'Ge-1 is a central component of the mammalian cytoplasmic mRNA processing body.', *Rna*, 11(12), pp. 1795–1802. doi: 10.1261/rna.2142405.

Zambon, R. A. *et al.* (2005) 'The Toll pathway is important for an antiviral response in Drosophila', *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 102(20), pp. 7257–7262. doi: 10.1073/pnas.0409181102.

Zehavi, Y. *et al.* (2014) 'The core promoter composition establishes a new dimension in developmental gene networks.', *Nucleus (Austin, Tex.)*. Taylor & Francis, 5(4), pp. 298–303. doi: 10.4161/nucl.29838.

Zelazny, B. (1973a) 'Studies on Rhabdionvirus oryctes: II. Effect on adults of Oryctes rhinoceros', *Journal of Invertebrate Pathology*. Academic Press, 22(1), pp. 122–126. doi:

10.1016/0022-2011(73)90020-7.

Zelazny, B. (1973b) 'Studies on Rhabdionvirus oryctes: III. Incidence in the Oryctes rhinoceros population of Western Samoa', *Journal of Invertebrate Pathology*. Academic Press, 22(3), pp. 359–363. doi: 10.1016/0022-2011(73)90164-X.

Zelazny, B. (1976) 'Transmission of a baculovirus in populations of Oryctes rhinoceros', *Journal of Invertebrate Pathology*. Academic Press, 27(2), pp. 221–227. doi: 10.1016/0022-2011(76)90149-X.

Zelazny, B. (1977) 'Oryctes rhinoceros populations and behavior influenced by a baculovirus', *Journal of Invertebrate Pathology*. Academic Press, 29(2), pp. 210–215. doi: 10.1016/0022-2011(77)90195-1.

Zerbino, D. R. and Birney, E. (2008) 'Velvet: algorithms for de novo short read assembly using de Bruijn graphs.', *Genome research*, 18(5), pp. 821–9. doi: 10.1101/gr.074492.107.

Zhao, H. H., Fernando, R. L. and Dekkers, J. C. M. (2007) 'Power and precision of alternate methods for linkage disequilibrium mapping of quantitative trait loci.', *Genetics*, 175(4), pp. 1975–86. doi: 10.1534/genetics.106.066480.

Zhao, J. *et al.* (2015) 'Recent advances on viral manipulation of NF-κB signaling pathway.', *Current opinion in virology*. NIH Public Access, 15, pp. 103–11. doi: 10.1016/j.coviro.2015.08.013.

Zhu, F., Ding, H. and Zhu, B. (2013) 'Transcriptional profiling of Drosophila S2 cells in early response to Drosophila C virus.', *Virology journal*, 10(1), p. 210. doi: 10.1186/1743-422X-10-210.

Zhu, F. and Zhang, X. (2013) 'The Wnt signaling pathway is involved in the regulation of phagocytosis of virus in Drosophila', *Scientific Reports*, 3(1), p. 2069. doi: 10.1038/srep02069.

Zuk, M. (2009) 'The sicker sex.', *PLoS pathogens*. Public Library of Science, 5(1), p. e1000267. doi: 10.1371/journal.ppat.1000267.

Appendix

Appendix to Chapter 2

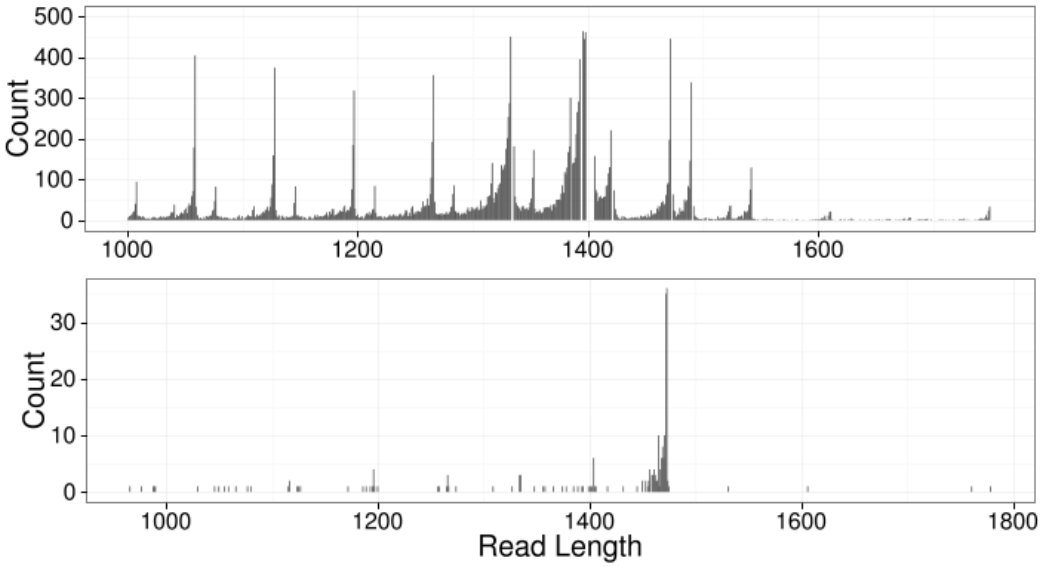

Figure S1: PacBio sequencing reads of the GRR in the DGRP

(Above) Read length distribution for all PacBio reads across GRR samples in the DGRP. Multiple peaks confirm that the population is variable in length for this region. (Below) An example DGRP line sequenced. Consensus sequences were created from reads which had a length supported by more than 10 reads.

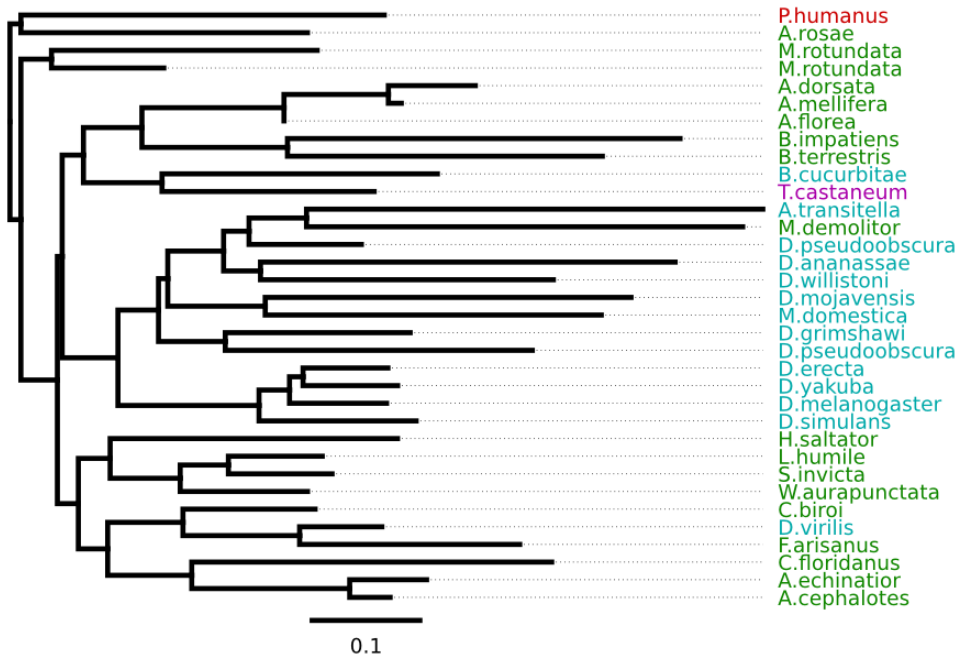

Figure S2: FFP profile clustering from repeat unit consensus sequence

The relationship between consensus sequences for GRR repeat units are unable to recover the true gene tree for Ago2 (consistent with bottom of Figure 1). Using only repeat consensus sequences should remove the effect of structure (which repeats are next to each other) on this similarity clustering, but reduces the amount of data.

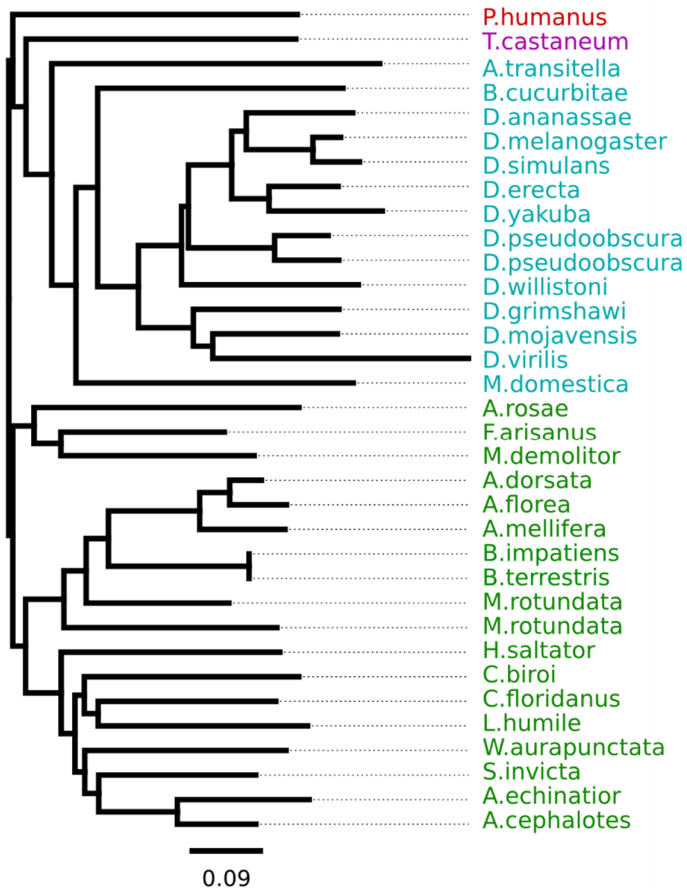

Figure S3: FFP profile clustering from conserved Ago2 sequence

FFP clustering on the conserved Ago2 sequence is mostly in accordance with the maximum clade credibility tree inferred using MrBayes (Figure 1, upper panel). This shows that the FFP clustering approach can infer an approximately correct gene tree from the remainder of Ago2, but not from the GRR region.

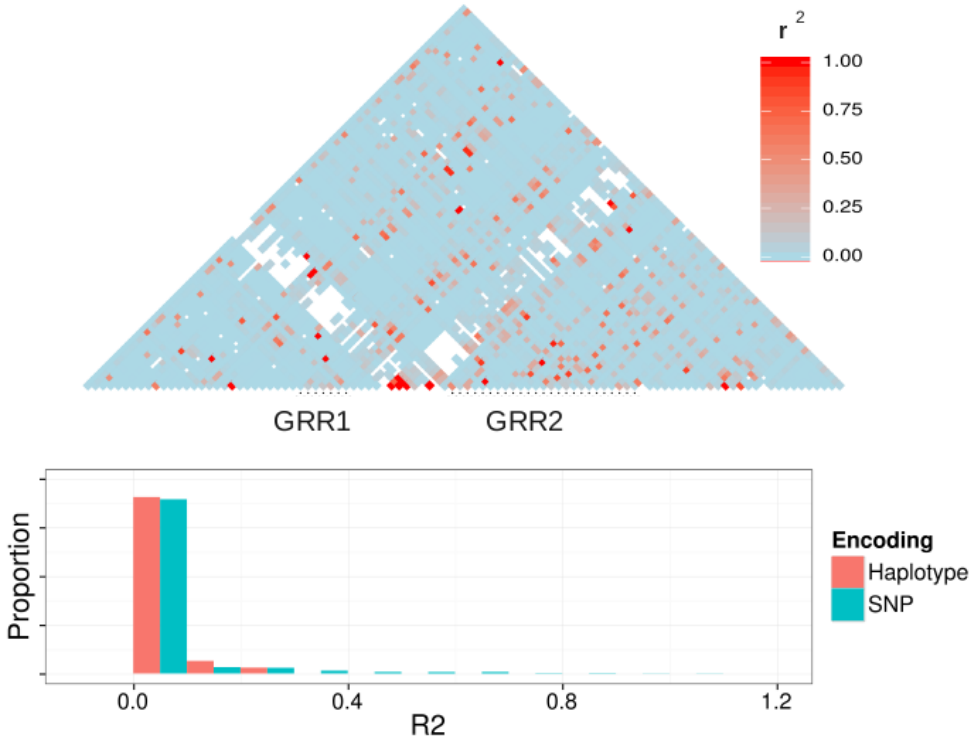

Figure S4: Linkage between GRR and surrounding 10KB and r^2 values between haplotypes and single GRR SNPs

Pairwise linkage diagram between GRR and surrounding area (above) and a histogram summarising these values. GRR alleles were either broken into a series of SNPs or compared as a single haplotype to calculate r^2. Linkage is overall low between this region and flanking area, and most information is not recovered by linked SNPs when haplotypes are used to calculate r^2. As expected, when coding the GRR as one highly multiallelic locus, LD with surrounding areas was much lower, as r^2 values for low frequency alleles is small and new haplotypes can be continually formed by both recombination and mutation.

Figure S5: Lower diversity in the GRR alpha group relative to GRR beta group

The log of the ratio of GRR alpha clade to GRR beta clade diversity is plotted (green line) as a sliding window along the chromosome at 50 kb to either side of Ago2. The diversity in haplotypes which carry the GRR alpha clade alleles show a surprising lack of diversity in a large area around Ago2. The red area shows the extremes of diversity differences expected by chance given the diversity in that region. The dotted lines show the gene region of Ago2.

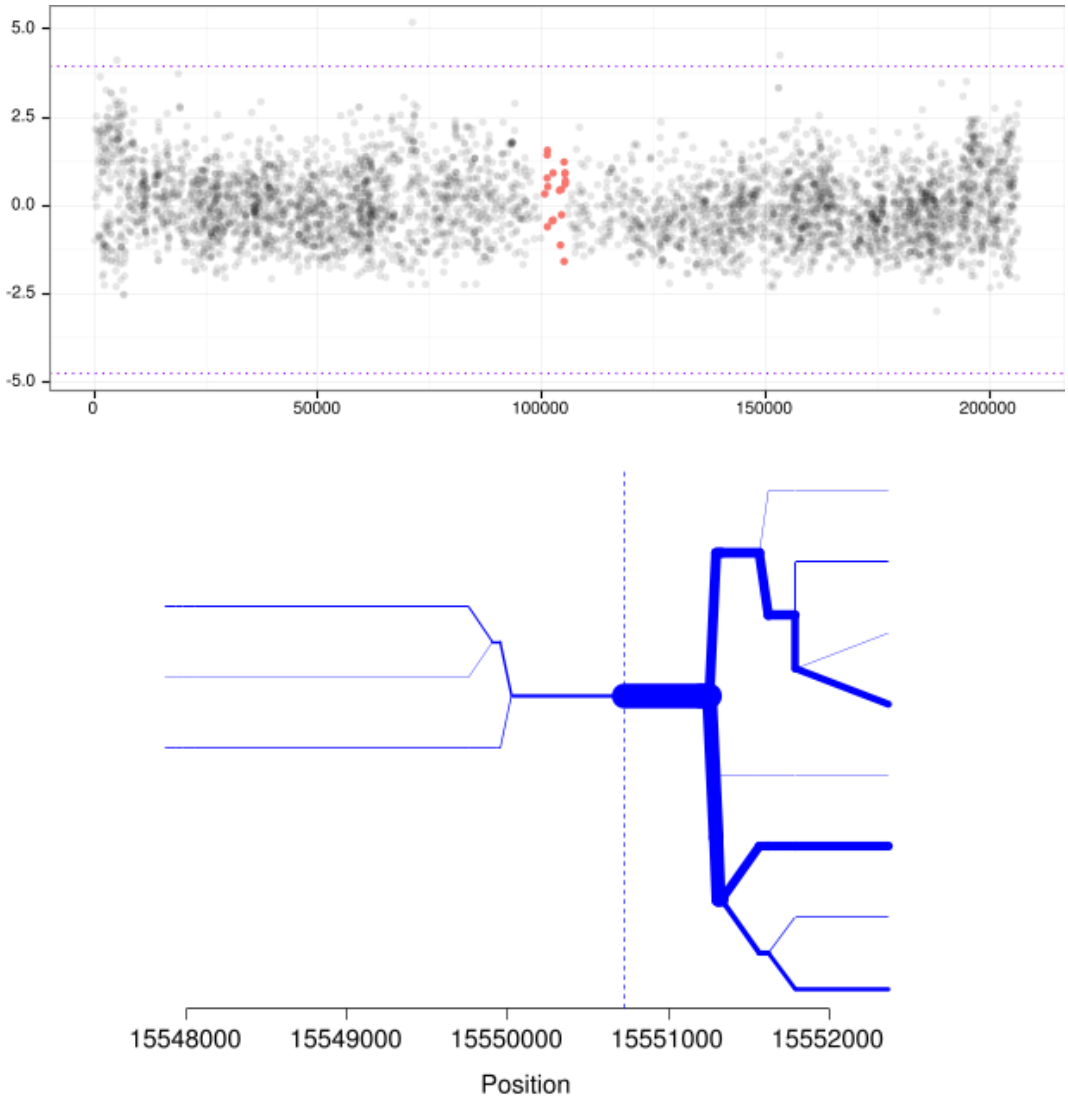

Figure S6: AGO2 nSL signature in the DGRP and haplotype bifurcation diagram

(Above) The frequency-standardised nSL statistic at each polymorphic site surrounding Ago2 indicates no abnormal haplotype structure (red points show polymorphic sites within the Ago2 gene, black points signify surrounding polymorphic sites). The dotted lines show a threshold at 3 interquartile ranges from the 1st and 3rd quartiles (i.e. extreme outliers). (Below) Visualising the breakdown of haplotype homozygosity supports this conclusion. The bifurcation diagram showing Extended Haplotype Homozygosity was created using the R package rehh (v.1.13) (Gautier and Vitalis, 2012). *Missing data for the intron upstream of the GRR2 in the DGRP means there is relatively few samples to calculate haplotype homozygosity for this region.*

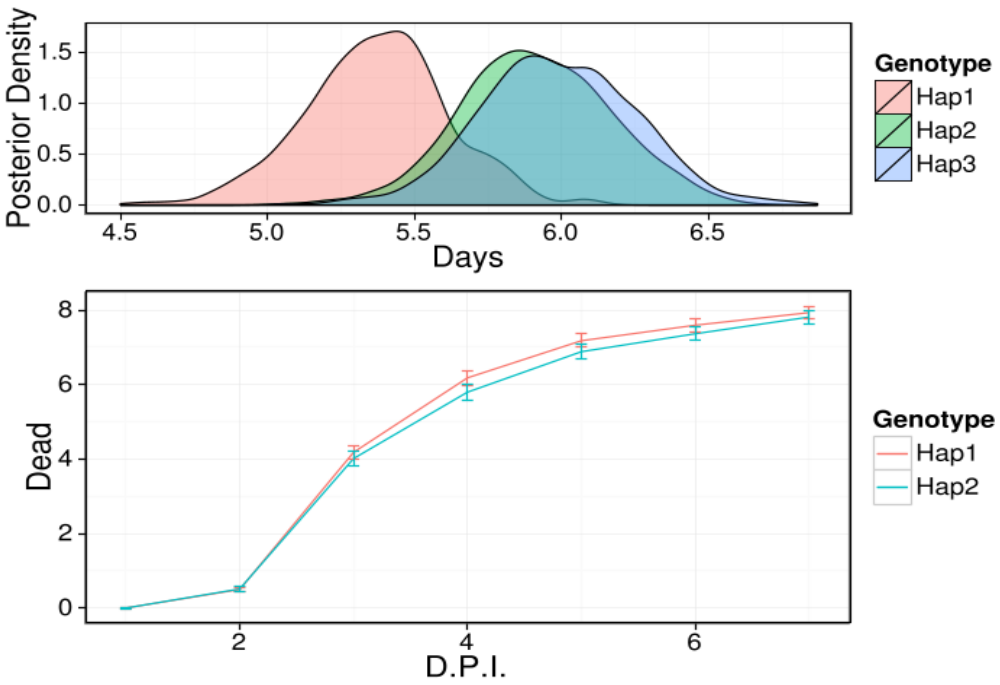

Figure S7: Involvement of GRR during DCV infection

(Upper) Estimated posterior density of the survival time (in days) post-infection for GRR Hap1, GRR Hap2 and GRR Hap3 using the GWAS data for the DGRP from Magwire et al (2012). GRR Hap2 and Hap3 are the two highest frequency haplotypes in the GRR beta clade (See Figure 2). (Below) However, upon infection with DCV, the two haplotype clades show no significant difference in mortality throughout infection in selected DSPR lines.

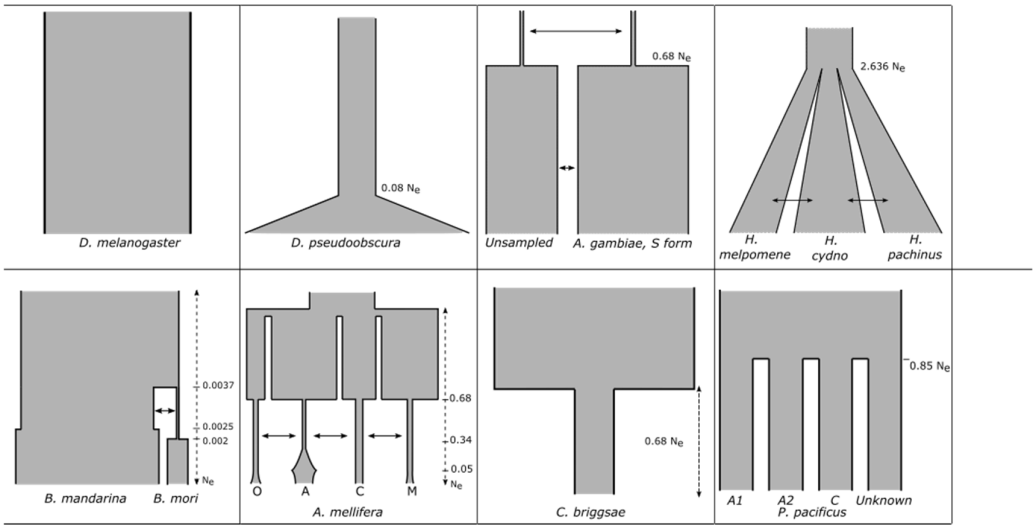

Figure S1 Demographic scenarios simulated for SweeD analysis

Coalescent simulations were performed using ms for demographic scenarios for each species which are supported by other studies. The African (Zambia) D. melanogaster were assumed to have a constant population size. D. pseudoobscura has recently undergone a population expansion 0.08 Ne generations ago. A. gambiae shares migrants with some other unknown, unsampled subpopulation which split 0.68 Ne generations ago. Heliconius species in Costa Rica split 2.636 Ne generations ago and have shared migrants since. Bombyx mandarina went through a small bottleneck when B. mori split, and shared migrants during that bottleneck (but not after). Apis mellifera have four subpopulations which have gone through multiple population expansions and bottlenecks, with all subpopulations sharing migrants until they join 0.68 Ne generations ago. Caenorhabditis briggsae "tropical samples" have undergone a population bottleneck 0.68 Ne generations ago. Finally, Pristionchus pacificus were sampled from four subpopulations, which split 0.85 Ne generations ago.

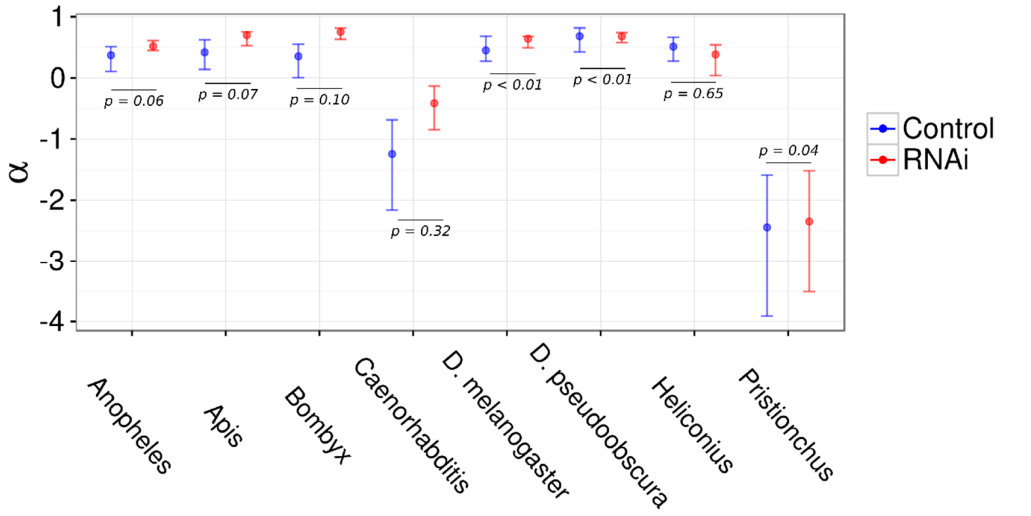

Figure S2: Alpha values for RNAi genes

For each species, α, or the proportion of adaptive substitutions was estimated from pooled polymorphism and divergence data using DFE-alpha for RNAi genes and position-matched control genes. α estimates for control genes are fairly constant across insect species, but are negative in the two nematode species. In all species except H. melpomene, the RNAi gene estimates are greater in RNAi genes than control genes.

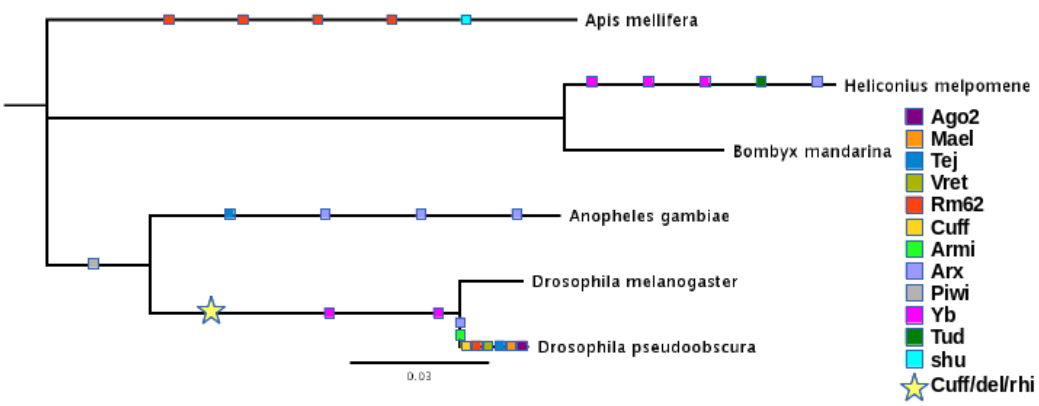

Figure S3 Possible duplications in RNAi pathway

Relationships of the insect species sampled, including coloured squares where possible gene duplications have occurred. Our search for RNAi genes in insect species other than D. melanogaster identified numerous duplications, and also some genes which were specific to Drosophila. Of note, D. pseudoobscura harboured duplications in asterix, armitage, cutoff, rm62, vretano, tejas, maelstrom, in addition to the multiple AGO2 duplications reported previously (Lewis et al, 2016; Lewis et al, 2016), perhaps indicating an extensive addition to RNAi related pathways. Asterix was further duplicated three times in Anopheles and once in Heliconius, and A. mellifera also has five distinct copies of rm62. Furthermore, yb duplications have occurred independently in the lineage leading to H. melpomene and the one leading to the Drosophila species. The piRNA cluster transcriptional complex composed of cutoff, deadlock, and rhino were only observed in the two Drosophila species (represented by a star), and thus have likely either been lost in the other species or have evolved in since the split between Anopheles and Drosophila.

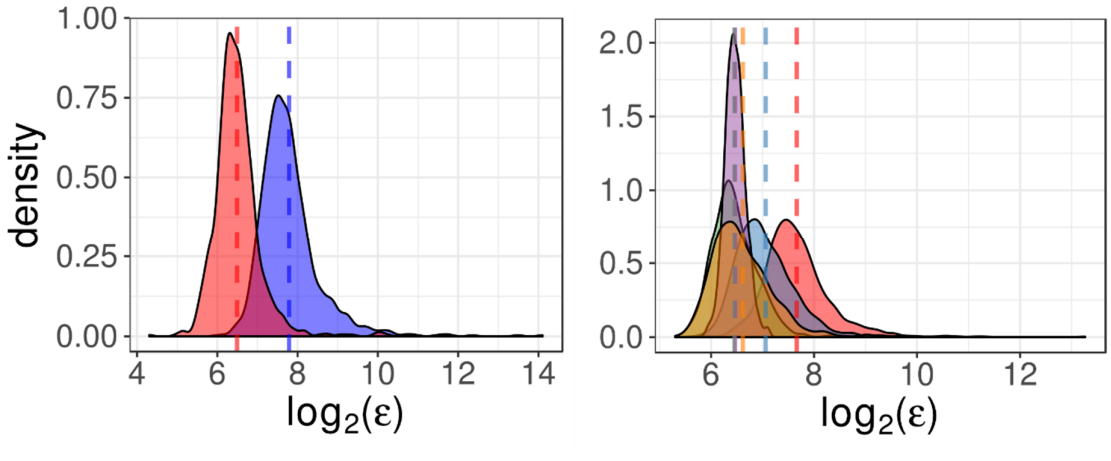

Figure S4 Coefficient of variation posterior distributions for residual variances in DFE-alpha meta-analysis

(Left) The coefficient of variation for the residual variance of RNAi genes (red) is not significantly greater than control genes (blue). (Right) This is also true for subpathway (piRNA: purple, viRNA: orange, miRNA: green, siRNA:blue) residual variances when combared with control genes (red). Therefore, the larger variances observed in rates of adaptive evolution in the RNAi pathway (Figure 1) and in RNAi subpathways (Figure 2) can be explained by an increase in the mean rates.

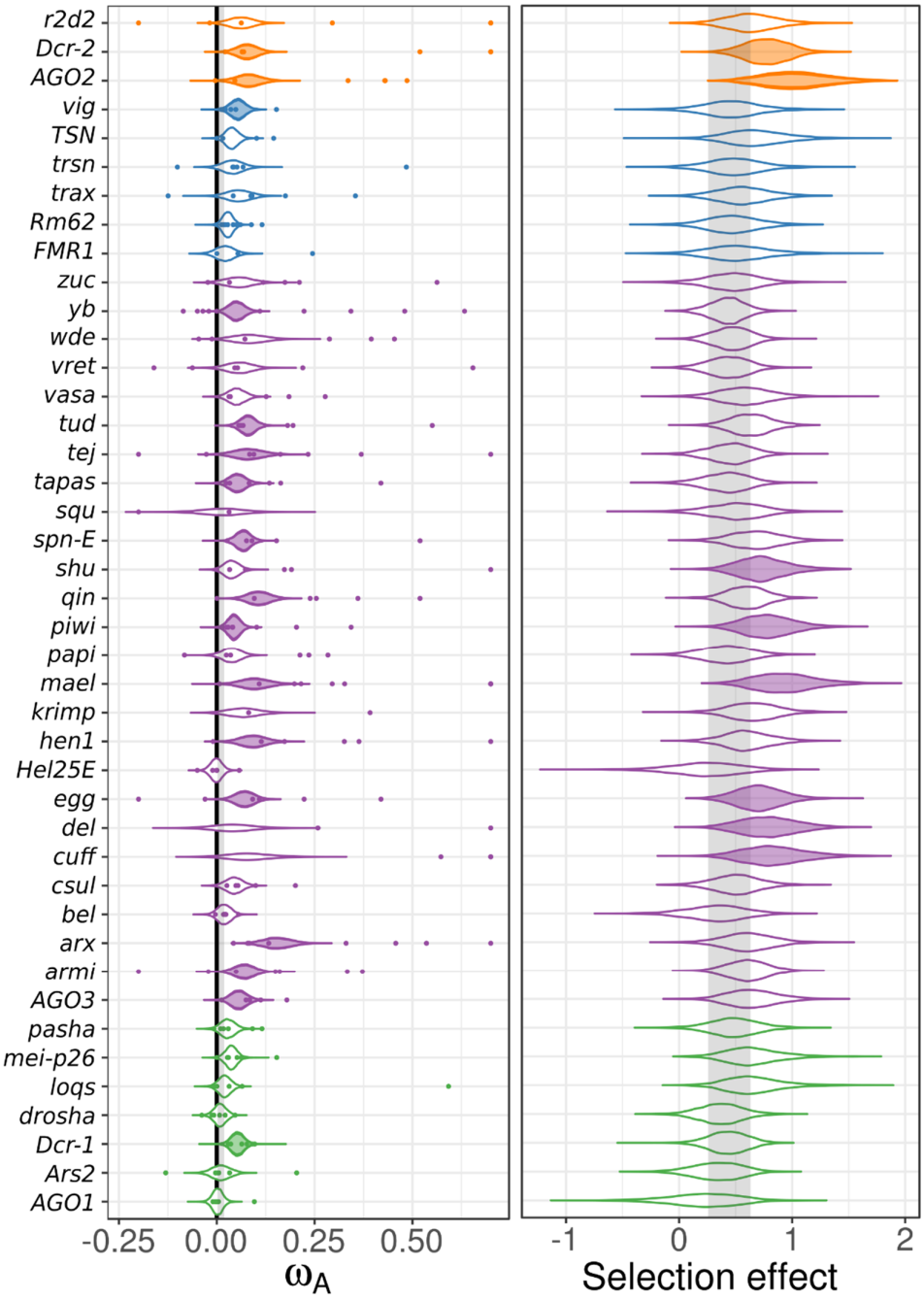

Figure S5 Cross-species homologue-level estimates of ω_A and selection effects without pathway assumptions

(Left) Individual gene ω_A estimates (coloured points) were calculated using DFE-alpha and analyses using a linear mixed model with species and gene as random effects (estimate uncertainty was included by incorporating bootstrap intervals as measurement error variance), but without subpathway as fixed effect (see Figure 4). The posterior distributions of the cross-species estimate for ω_A for each gene are plotted, and shaded if the MCMCp < 0.05 when tested against the control gene distribution (shaded grey region). Single-gene estimates of $\omega_A > 0.75$ are plotted at 0.75 for clarity. (Right) The analogous analysis, except performed using SnIPRE, with the posterior distribution of homologue-level selection effects plotted. Both analyses find AGO2, Dicer-2, piwi, maelstrom, and eggless as having elevated protein substitution.

Species	Gene	Duplicate	Selection Effect	95% HPD interval	Species	Gene	Duplicate	Selection Effect	95% HPD interval
D. melanogaster	arx	A	1.16	[0.157-2.09]	*D. pseudoobscura*	vas	A	1.16	[0.28-2.04]
D. melanogaster	piwi	A	1.11	[0.294-1.94]	*D. pseudoobscura*	vig	A	1.49	[0.337-2.65]
D. melanogaster	csul	A	1.53	[0.693-2.33]	*D. pseudoobscura*	vret	B	1.48	[0.703-2.3]
D. melanogaster	Dcr-1	A	1.21	[0.441-2]	*D. pseudoobscura*	wde	A	0.876	[0.251-1.5]
D. melanogaster	Dcr-2	A	1.03	[0.151-1.89]	*D. pseudoobscura*	Yb	A	1.75	[0.871-2.74]
D. melanogaster	krimp	A	1.12	[0.174-1.94]	*D. pseudoobscura*	Yb	B	0.737	[0.144-1.31]
D. melanogaster	qin	A	1.04	[0.421-1.61]	*D. pseudoobscura*	zuc	A	1.43	[0.649-2.15]
D. melanogaster	r2d2	A	1.59	[0.625-2.66]	*Heliconius*	shu	A	0.756	[0.0191-1.51]
D. melanogaster	tapas	A	1.35	[0.639-2.03]	*Heliconius*	Yb	A	1.03	[0.483-1.57]
D. melanogaster	tej	A	0.826	[0.21-1.4]	*Apis*	AGO2	A	1.55	[0.76-2.37]
D. melanogaster	tud	A	0.816	[0.0691-1.51]	*Apis*	piwi	A	1.04	[0.306-1.79]
D. melanogaster	vas	A	1.63	[0.975-2.33]	*Apis*	Dcr-2	A	0.965	[0.195-1.72]
D. melanogaster	vret	A	0.995	[0.211-1.72]	*Apis*	rm62	C	1.12	[0.02-2.13]
D. melanogaster	Yb	C	1.2	[0.472-1.93]	*Apis*	tud	A	0.941	[0.253-1.61]
D. melanogaster	SoYb	B	0.78	[0.262-1.3]	*Apis*	vas	A	1.46	[0.467-2.48]
D. pseudoobscura	AGO2	A	1.07	[0.307-1.86]	*Anopheles*	AGO3	A	0.934	[0.103-1.85]
D. pseudoobscura	armi	B	2.66	[2.09-3.25]	*Anopheles*	armi	A	0.82	[0.0614-1.57]
D. pseudoobscura	armi	A	1.22	[0.622-1.82]	*Anopheles*	arx	A	1.39	[0.376-2.33]
D. pseudoobscura	arx	B	1.01	[0.0294-2.11]	*Anopheles*	piwi	B	1.26	[0.569-1.92]
D. pseudoobscura	piwi	A	0.952	[0.043-1.81]	*Anopheles*	csul	A	1.04	[0.141-1.91]
D. pseudoobscura	csul	A	0.921	[0.00777-1.83]	*Anopheles*	Dcr-1	A	0.772	[0.101-1.43]
D. pseudoobscura	cuff	A	1.26	[0.41-2.12]	*Anopheles*	drosha	A	1.54	[0.721-2.37]
D. pseudoobscura	cuff	B	1.33	[0.458-2.16]	*Anopheles*	loqs	A	1.68	[0.794-2.54]
D. pseudoobscura	Dcr-2	A	1.48	[0.762-2.27]	*Anopheles*	mei-p26	A	1.08	[0.234-1.89]
D. pseudoobscura	del	A	1.33	[0.413-2.17]	*Anopheles*	papi	A	0.98	[0.203-1.79]
D. pseudoobscura	drosha	A	1.62	[0.788-2.47]	*Anopheles*	tej	A	0.868	[0.169-1.53]
D. pseudoobscura	egg	A	0.949	[0.249-1.63]	*Anopheles*	trax	A	1.48	[0.433-2.48]
D. pseudoobscura	loqs	A	1.33	[0.301-2.42]	*Anopheles*	tud	A	1.37	[0.716-2.03]
D. pseudoobscura	mael	A	1.22	[0.0958-2.36]	*Anopheles*	vig	A	1.29	[0.196-2.4]
D. pseudoobscura	mael	B	1.22	[0.0897-2.36]	*Anopheles*	wde	A	0.957	[0.338-1.58]
D. pseudoobscura	mei-p26	A	1.66	[0.852-2.38]	*Anopheles*	zuc	A	1.42	[0.535-2.31]
D. pseudoobscura	papi	A	0.79	[0.0114-1.56]	*Bombyx*	AGO2	A	1.41	[0.633-2.18]
D. pseudoobscura	pasha	A	1.2	[0.247-2.14]	*Bombyx*	Dcr-1	A	2.95	[2.14-3.75]
D. pseudoobscura	qin	A	1.56	[0.975-2.16]	*Bombyx*	egg	A	2.06	[1.26-2.92]
D. pseudoobscura	r2d2	A	1.29	[0.22-2.37]	*Bombyx*	Hel25E	A	1.18	[0.0422-2.35]
D. pseudoobscura	rm62	B	1.33	[0.247-2.29]	*Bombyx*	mael	A	1.04	[0.324-1.75]
D. pseudoobscura	rm62	A	1.09	[0.0839-2.01]	*Bombyx*	papi	A	1.15	[0.11-2.18]
D. pseudoobscura	shu	A	1.57	[0.667-2.48]	*Bombyx*	shu	A	1.31	[0.64-1.97]
D. pseudoobscura	spn-E	A	1.97	[1.31-2.59]	*Bombyx*	tapas	A	0.989	[0.233-1.74]
D. pseudoobscura	squ	A	1.19	[0.26-2.09]	*Bombyx*	trsn	A	1.73	[0.595-2.94]
D. pseudoobscura	tej	A	1.13	[0.171-2]	*Bombyx*	vig	A	1.72	[0.432-2.92]
D. pseudoobscura	trsn	A	1.54	[0.622-2.56]	*Bombyx*	vret	A	0.714	[0.0928-1.33]
D. pseudoobscura	tud	A	0.912	[0.343-1.53]	*Bombyx*	Yb	A	1.34	[0.778-1.87]

Figure S6 Genes with elevated selection effects

A list of RNAi genes with individually significant selection effects (i.e. a selection effect > 0) and their 95% highest posterior density intervals.

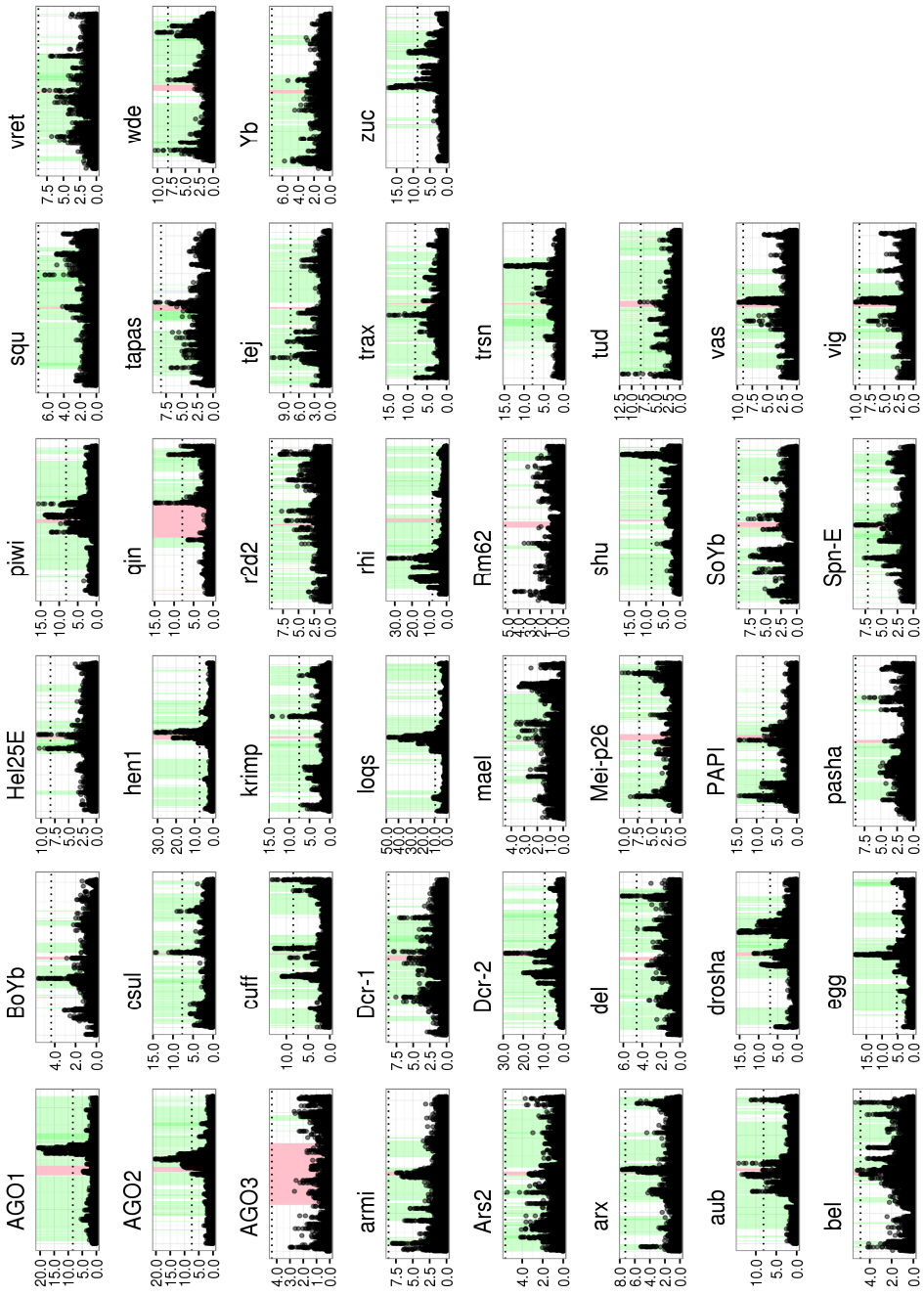

Figure S7: Drosophila melanogaster sweeps

For each D. melanogaster gene, the CLR statistic was plotted across a 200 kb region including the gene of interest. Each panel represents a region of the D. melanogaster genome, with red-shaded regions being the gene of interest and green-shaded regions being other genes along the chromosome. The horizontal dotted lines in each panel are significance thresholds calculated through neutral coalescent simulations.

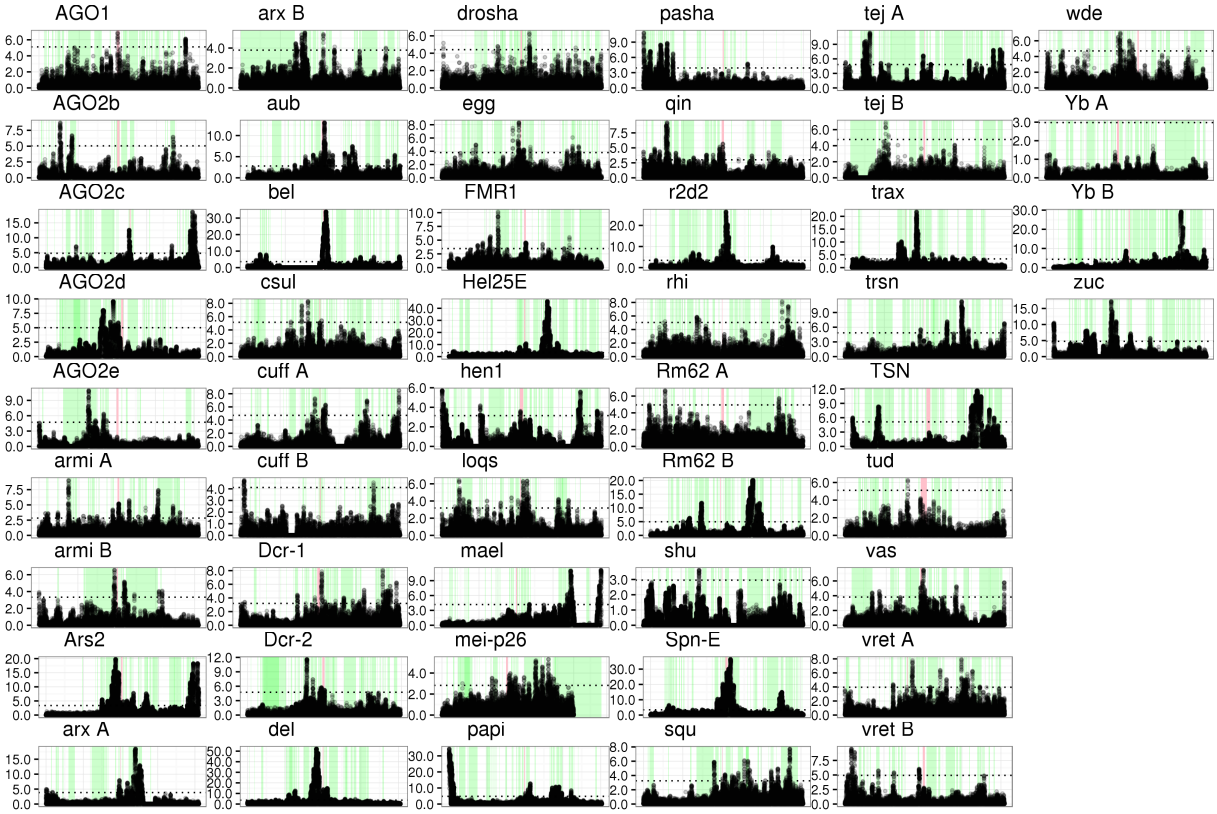

Figure S8: Drosophila pseudoobscura sweeps

For each D. pseudoobscura gene, the CLR statistic was plotted across a 200 kb region including the gene of interest. Each panel represents a region of the D. pseudoobscura genome, with red-shaded regions being the gene of interest and green-shaded regions being other genes along the chromosome. The horizontal dotted lines in each panel are significance thresholds calculated through neutral coalescent simulations.

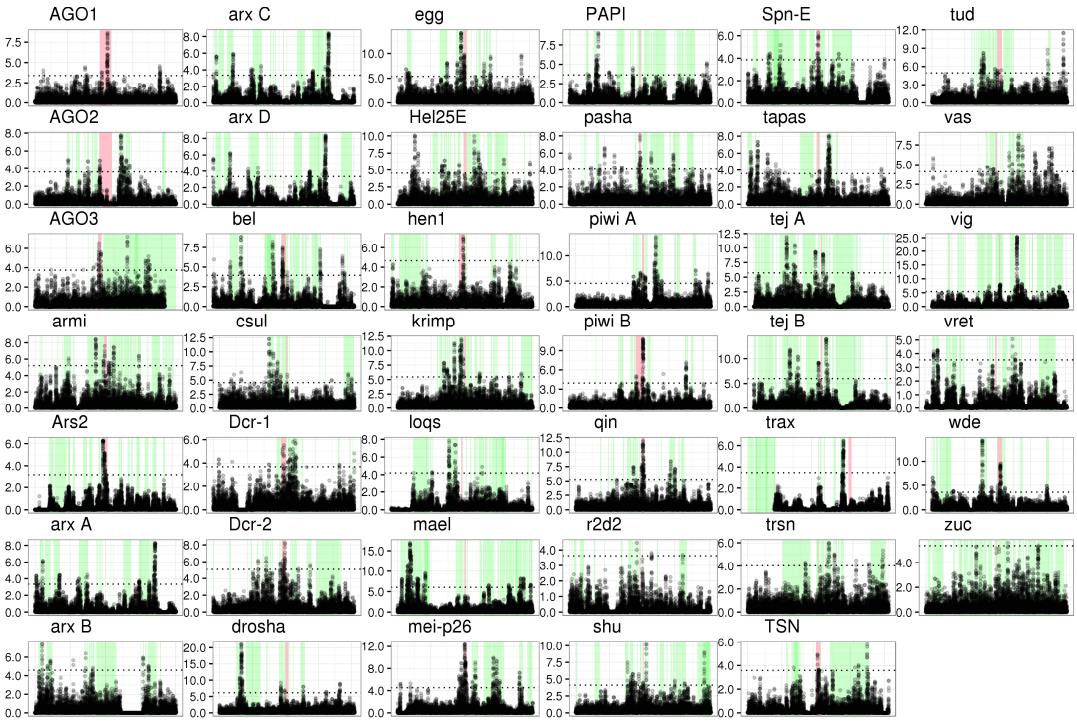

Figure S9: Anopheles gambiae sweeps

For each A. gambiae gene, the CLR statistic was plotted across a 200 kb region including the gene of interest. Each panel represents a region of the A. gambiae genome, with red-shaded regions being the gene of interest and green-shaded regions being other genes along the chromosome. The horizontal dotted lines in each panel are significance thresholds calculated through neutral coalescent simulations.

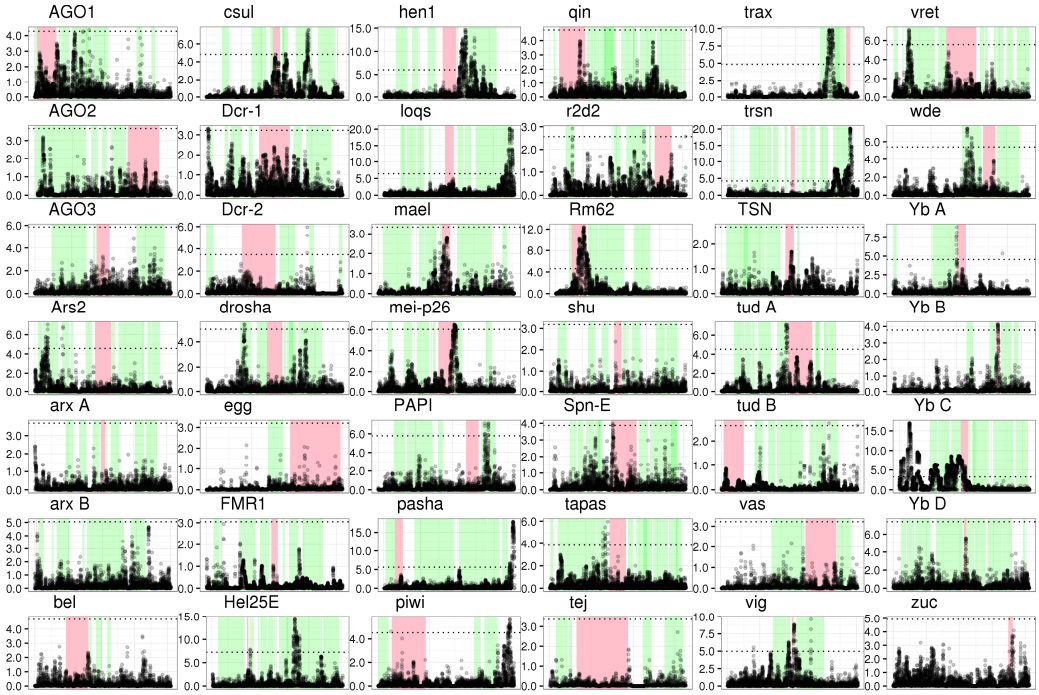

Figure S10: Heliconius melpomene sweeps

For each H. melpomene gene, the CLR statistic was plotted across a 200 kb region including the gene of interest. Each panel represents a region of the H. melpomene genome, with red-shaded regions being the gene of interest and green-shaded regions being other genes along the chromosome. The horizontal dotted lines in each panel are significance thresholds calculated through neutral coalescent simulations.

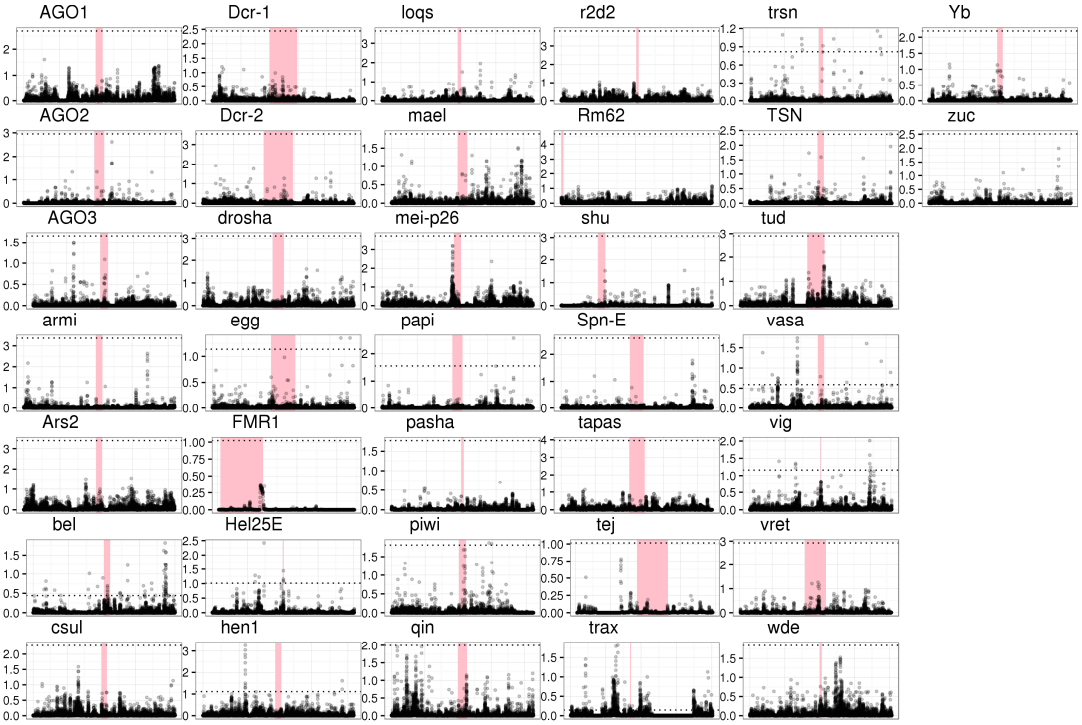

Figure S11: Bombyx mandarina sweeps

For each B. mandarina gene, the CLR statistic was plotted across a 200 kb region including the gene of interest. Each panel represents a region of the B. mandarina genome, with red-shaded regions being the gene of interest. The horizontal dotted lines in each panel are significance thresholds calculated through neutral coalescent simulations. The Bombyx genome used did not have an associated gff file, and so positions of nearby genes were not included.

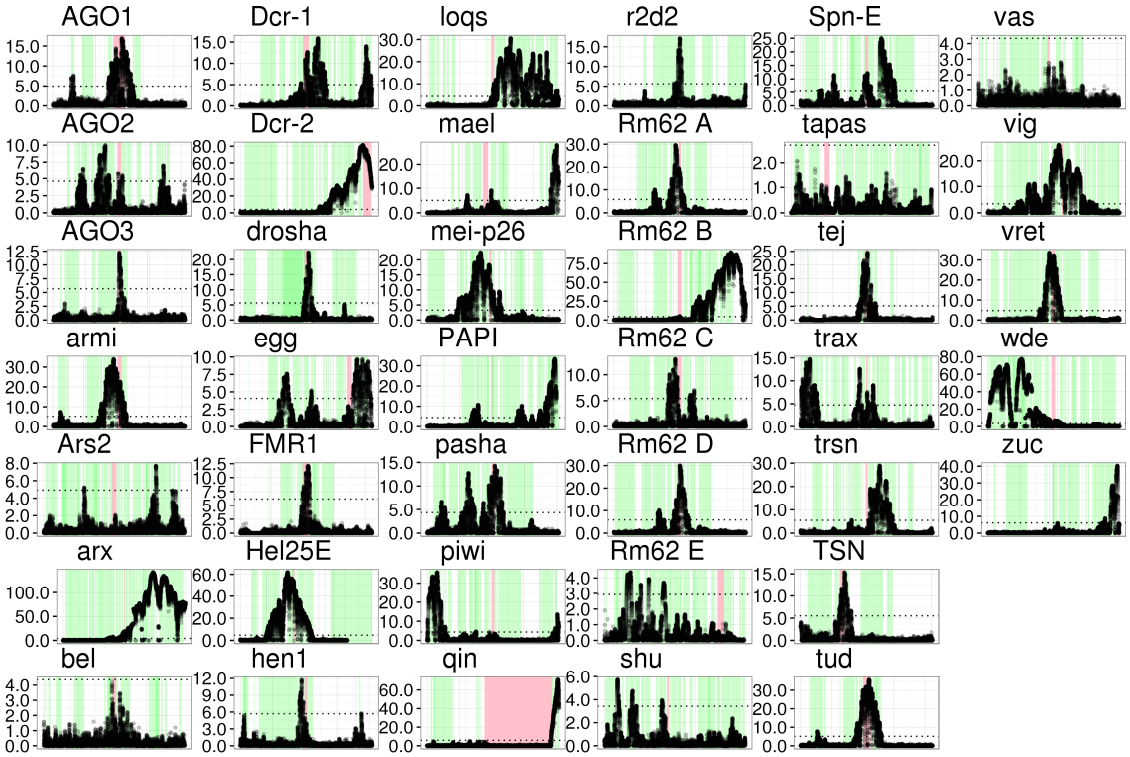

Figure S12: Apis mellifera sweeps

For each A. mellifera gene, the CLR statistic was plotted across a 200 kb region including the gene of interest. Each panel represents a region of the A. mellifera genome, with red-shaded regions being the gene of interest. The horizontal dotted lines in each panel are significance thresholds calculated through neutral coalescent simulations.

Figure S13: Caenorhabditis briggsae sweeps

For each C. briggsae gene, the CLR statistic was plotted across a 200 kb region including the gene of interest. Each panel represents a region of the C. briggsae genome, with red-shaded regions being the gene of interest. The horizontal dotted lines in each panel are significance thresholds calculated through neutral coalescent simulations.

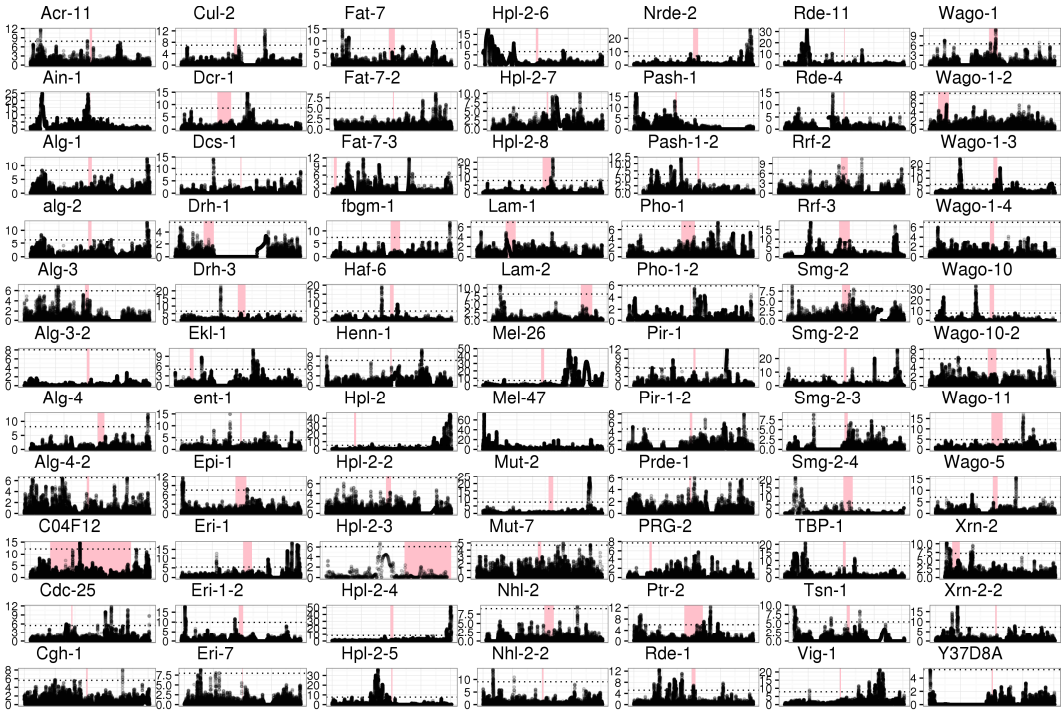

Figure S14: Pristionchus pacificus sweeps

For each P. pacificus gene, the CLR statistic was plotted across a 200 kb region including the gene of interest. Each panel represents a region of the P. pacificus genome, with red-shaded regions being the gene of interest. The horizontal dotted lines in each panel are significance thresholds calculated through neutral coalescent simulations. The Pristionchus genome used did not have an associated gff file, so positions of nearby genes were not included.

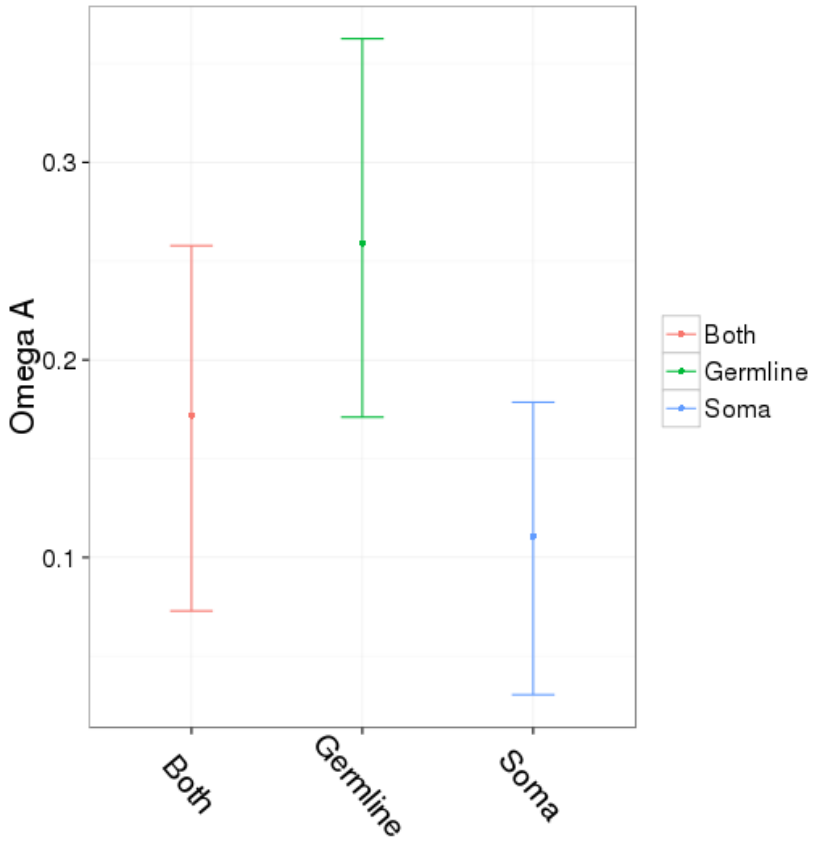

Figure S15: Germline and somatic piRNA pathway genes

Polymorphism and divergence from a larger set of piRNA pathway genes (those identified in two of three recent piRNA pathway screens, plus the core piRNA pathway) (Handler et al, 2013; Czech et al, 2013; Muerdter et al, 2013) in D. melanogaster were pooled based on whether they are active in the germline, soma, or both and used to calculate ω_A. Confidence intervals were obtained by bootstrapping by gene 1000 times. Genes active in germline TE suppression show higher rates of adaptive protein evolution than those active in the somatic follicle cells, with genes active in both having an intermediate adaptive rate.

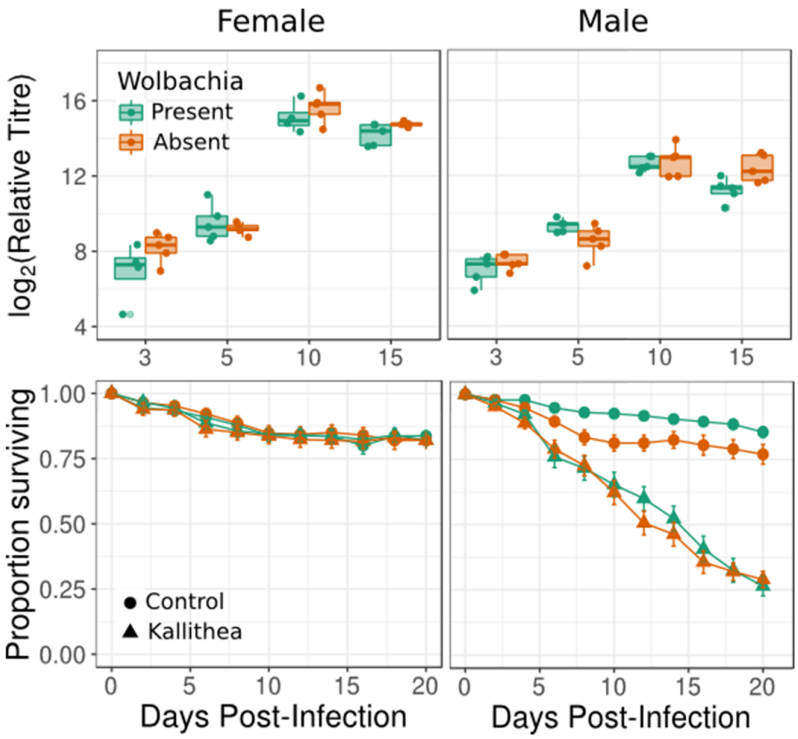

Figure S1: No effect of Wolbachia on KV growth or KV-induced mortality

Upper panels: Log-transformed relative viral titre in Wolbachia positive (green) or negative (orange) OreR female and male flies. Lower panels: mortality curves for gradient control-injected (circle) or KV-injected (triangle) OreR female and male flies, with or without Wolbachia.

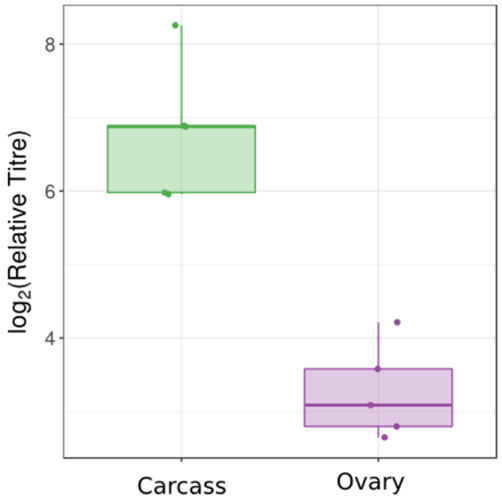

Figure S2: KV enriched in carcass relative to ovary

Females had higher viral titres in non-ovary tissues at 3 DPI. However, this could be affected by the route of infection, and the average ploidy of each sample.

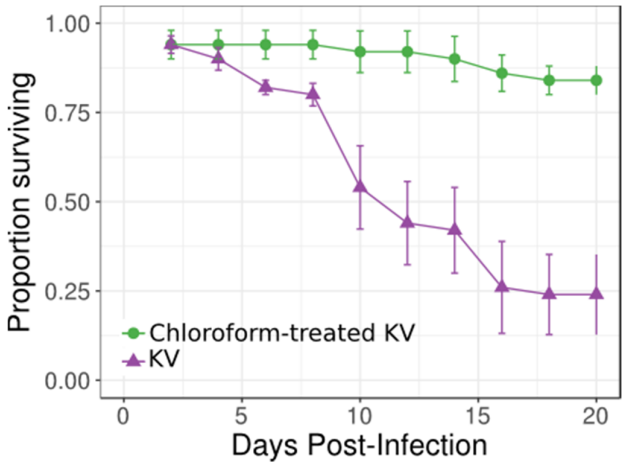

Figure S3: KV-induced male mortality is not caused by contaminating RNA viruses

Chloroform-treatment is expected to inactivate enveloped viruses such as KV, but unenveloped viruses (including most +ssRNA viruses) are expected to retain infectivity. We confirmed mortality following KV infection was not caused by contaminating DAV by comparing injection of the KV isolate with (green) or without (purple) inactivating chloroform treatment.

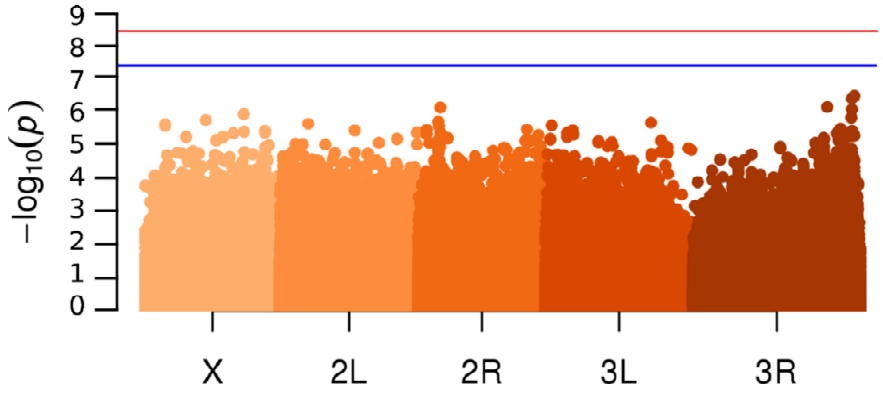

Figure S4: Sex-by-genotype interaction significance manhattan plot

No polymorphism had a significant effect on sex-specific mortality. The blue line denotes p_{rand} = 0.05 and the red line is p_{rand} = 0.01.

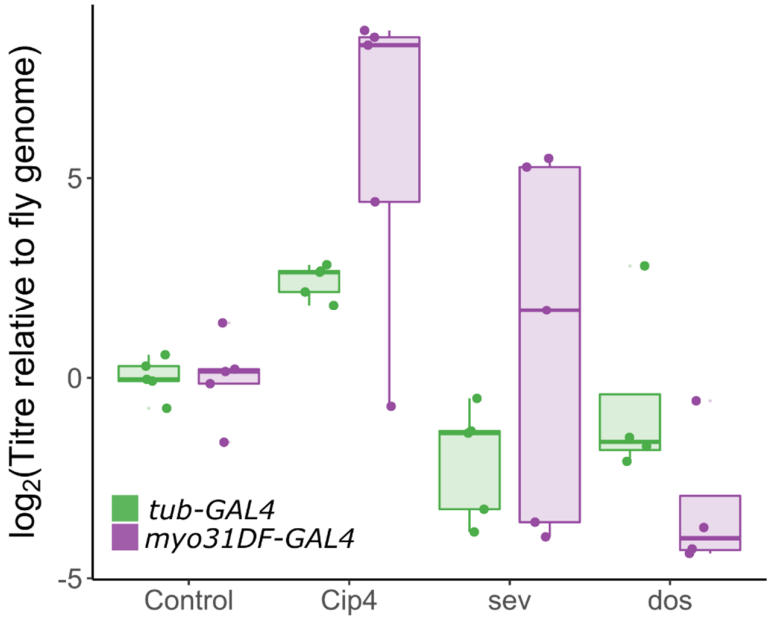

Figure S5: Virus titre in other confirmed GWAS knock-down lines

KV titre was measured in flies expressing a foldback hairpin targeting 18 genes identified in the GWAS, using GAL4 lines that knock each down in either the whole fly (tub-GAL4, green) or specifically in the gut (myo31DF-GAL4, purple). Only those causing a significant increase in titre relative to other knock-down lines (e.g. Figure 5) are shown here. Note that the cross between myo31DF-GAL4 and CG12821^{IR} was inexplicably lethal, and that titre was highly variable in some of the other myo31DF-GAL4 crosses.

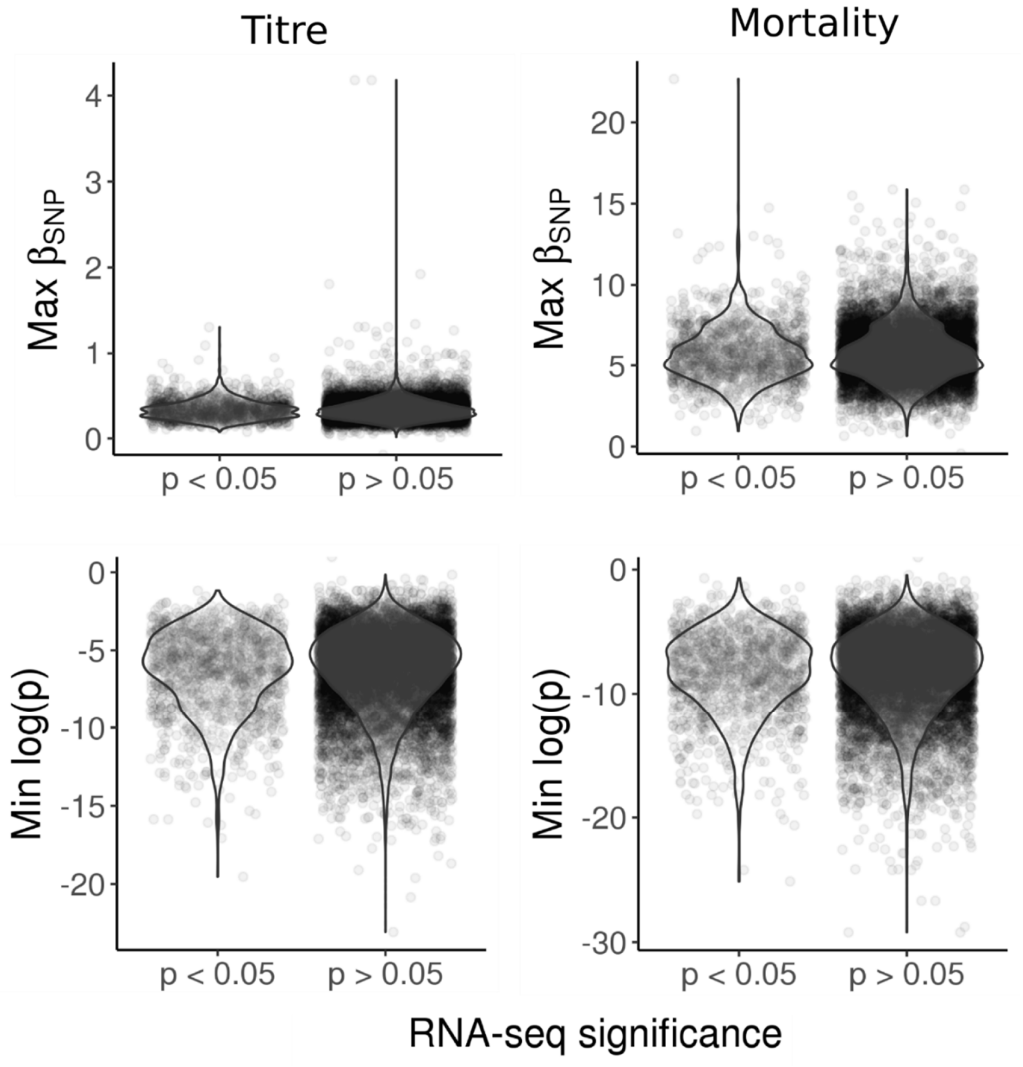

Figure S6: No enrichment of KV infection-associated SNPs in KV-responsive genes

Genes were split into KV-responsive and KV-unresponsive genes based on the RNA sequencing differential expression analysis ($p < 0.05$). The largest effect size (max β^{SNP}) and lowest p-value was recorded for each gene in each GWAS, and compared between the KV-responsive and unresponsive genes. We find no significant difference between any comparisons.

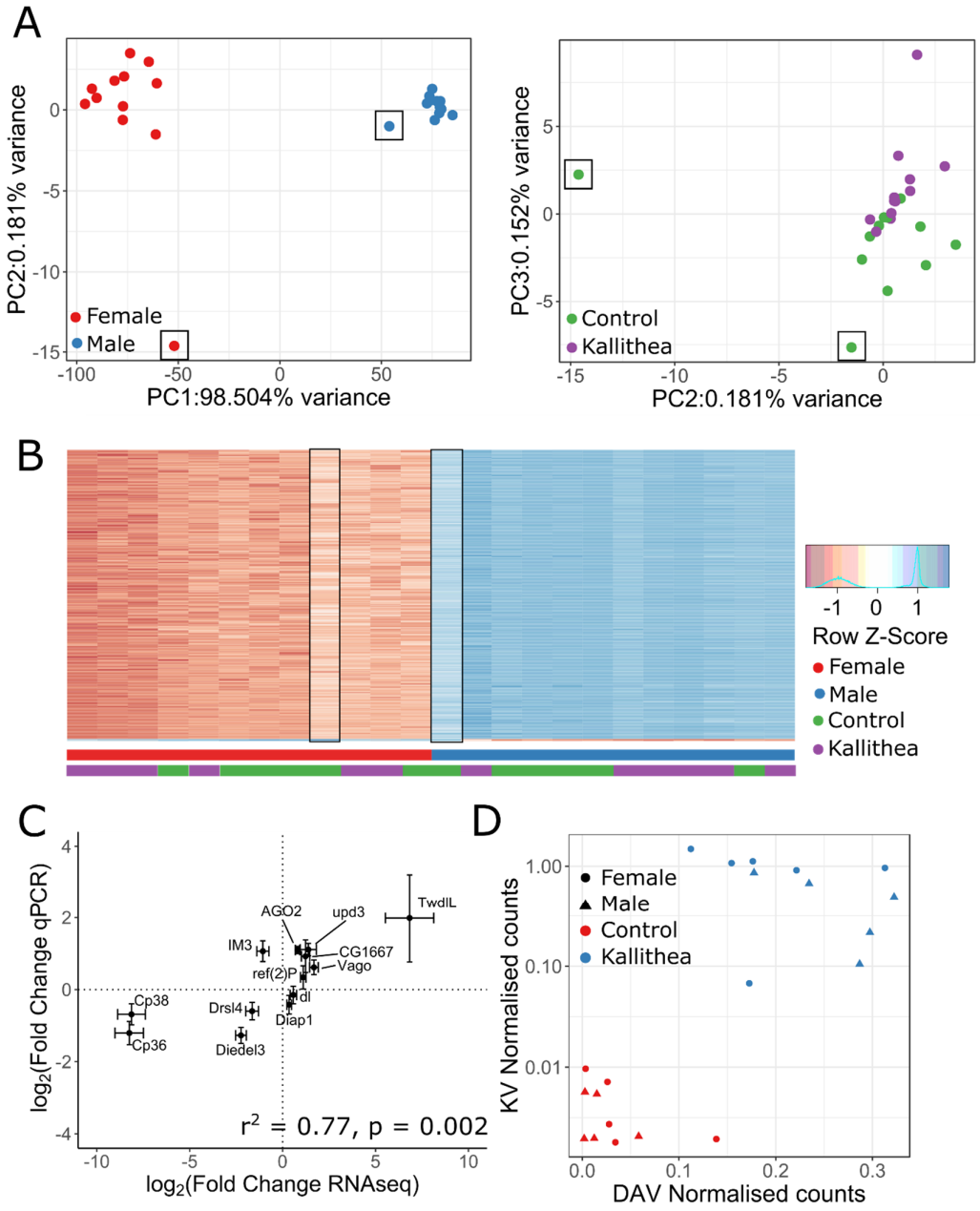

Figure S7: RNA sequencing: Library quality and verification

(A) The first three principal components of read counts per gene in RNA-sequencing data, plotted such that each library is represented by a point. Males (blue) and females (red) are separated on PC1. Control-injected (green) and KV-injected (purple) are separated on PC3. For (B), we clustered libraries based on expression of the 1000 most variable genes, where each row on the heatmap is a gene, and the columns are libraries. Together, these analyses identified two possible outlier libraries, which were excluded (black rectangles in A and B). (C) We selected 13 well-studied immune genes and genes with a clear phenotype association (e.g. chorion proteins), distributed across the range of differential expression values, for qPCR verification. Using 5 independent biological replicates from the outbred DGRP population, we confirmed that differential expression for these genes was highly correlated between qPCR and RNA-seq ($r^2 = 0.77$, $p = 0.002$). (D) We found low-level DAV contamination in our RNA-sequencing experiment. The plot shows the relationship between DAV viral titre and average KV gene expression, where each point is the number of reads mapping to KV and DAV for each library, normalised by library size factor and genome length. Note that KV is plotted on a $\log_{10}$ scale, but DAV on a linear scale.

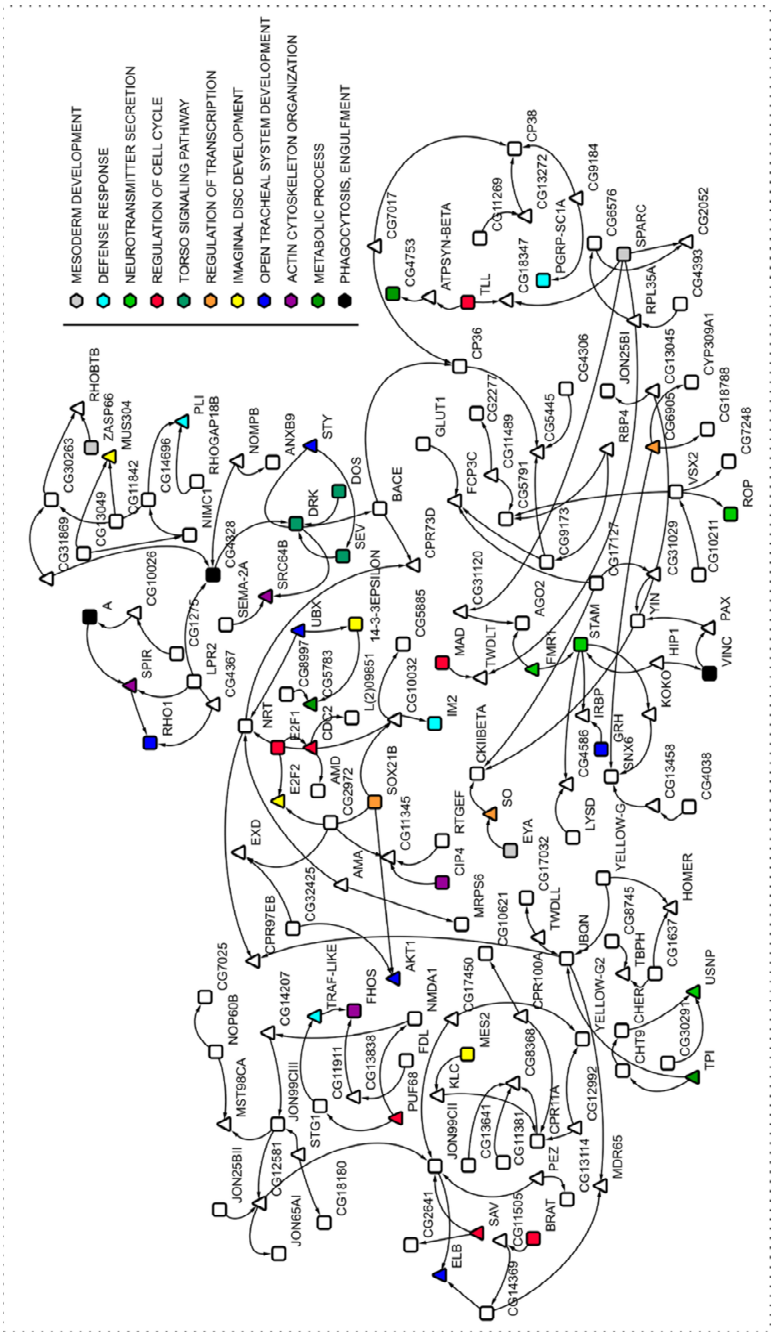

Figure S8: Pathway analysis of genes identified by differential expression analysis or GWAS

Network analysis identified a pathway marginally enriched for genes identified in these studies (p = 0.03). Shown are interactions (edges) between genes identified in GWAS or RNA-sequencing studies (square nodes), allowing a single gene not included in the original gene list to mediate an interaction (triangular nodes). Genes (nodes) are coloured by functional annotation and white nodes are unannotated. Node labels are associated with the bottom-right corners of nodes.

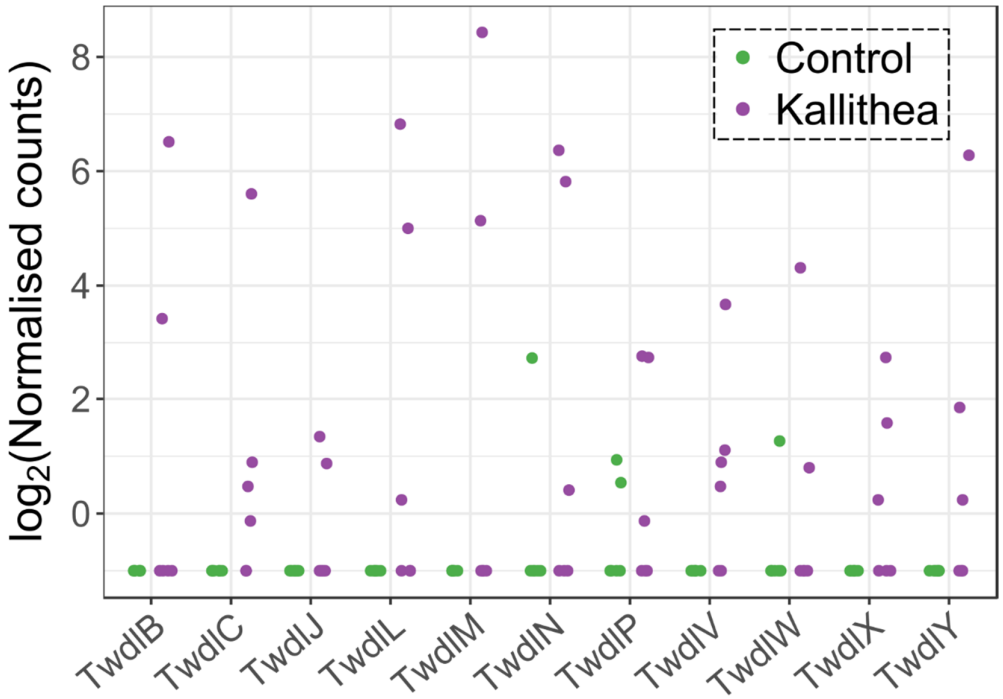

Figure S9: Variable differential expression of the Tweedle gene family

A subset of KV-infected (purple) vials showed very high expression of Tweedle genes, whereas these were mostly unexpressed in control (green) adult flies.

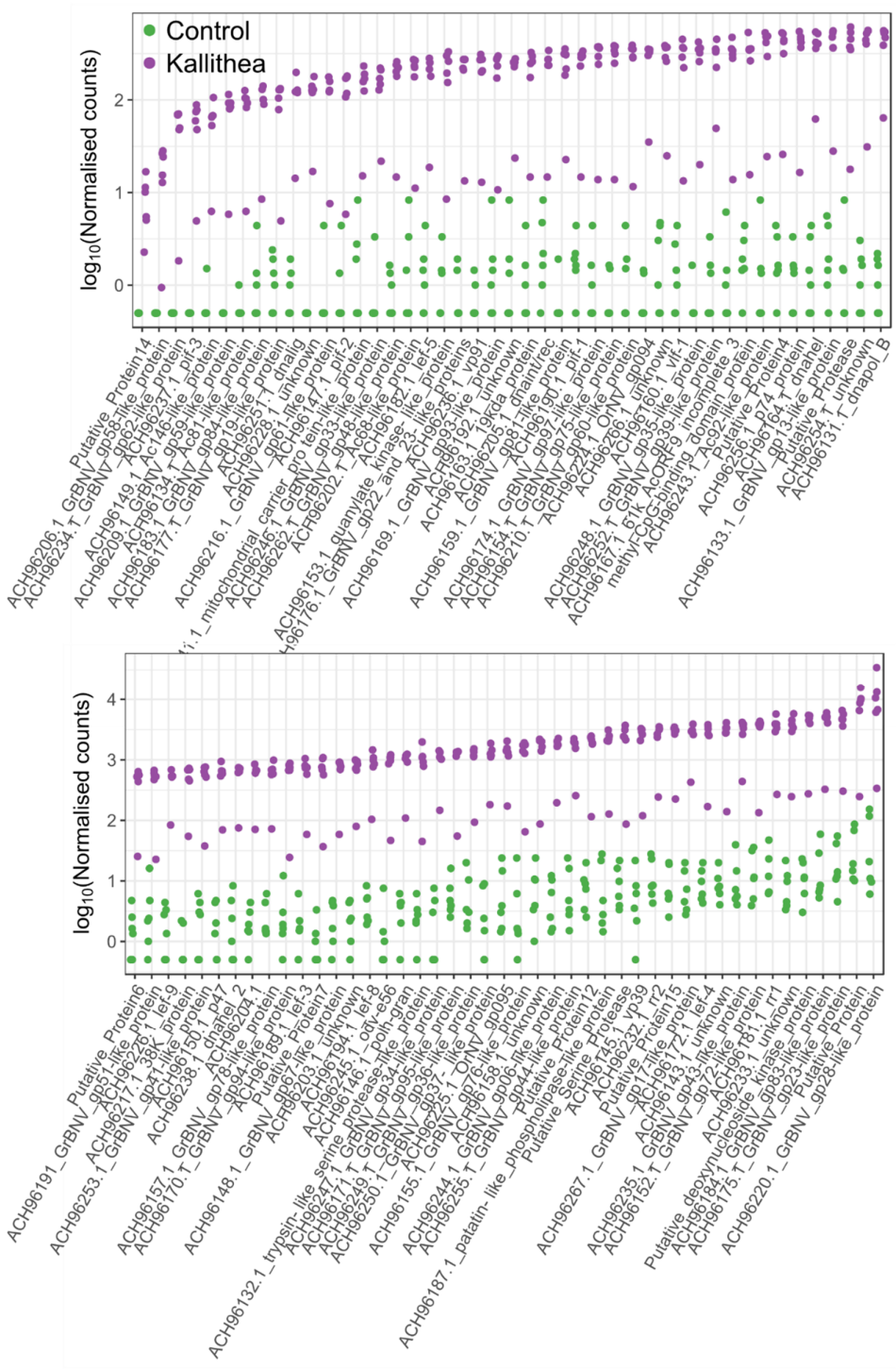

Figure S10: Expression of Kallithea Virus genes

Most KV genes are expressed at 3 DPI. One KV-injected vial of flies had a lower level of infection. Control libraries also showed mapping to KV genes, most likely due to a low level (<0.5%) of barcode switching among libraries run together. The lower panel is a continuation of the upper panel.

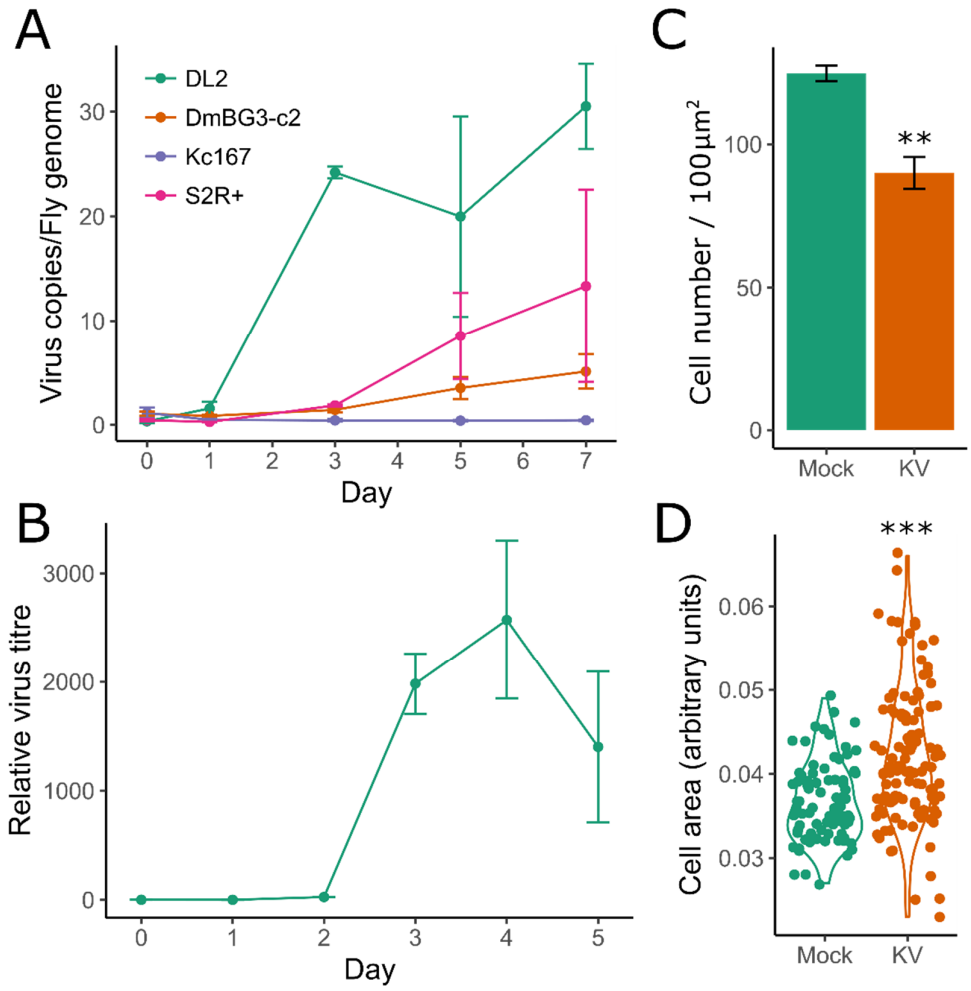

Figure S1: KV growth dynamics in cell culture

KV had different growth rates in different cell lines (A), and grew best in DL2 and S2R+ cells (n=3 for each time point). (B) We performed qPCR on cell culture medium of KV-infected cells, and found KV was released from cells starting from 3 DPI (n = 3). Although we did not observe any overt cytopathic effects of KV-infected cells, passage of cells 7 DPI revealed a slower growth rate (C), evidenced by the number of cells per approximately 100 µm² in KV versus mock-treated cells (n=3). (D) The passaged KV-infected cells also had a subset of significantly larger cells. Error bars show standard error of the mean. Experiments for (B,C,D) were performed with S2 cells from Invitrogen.

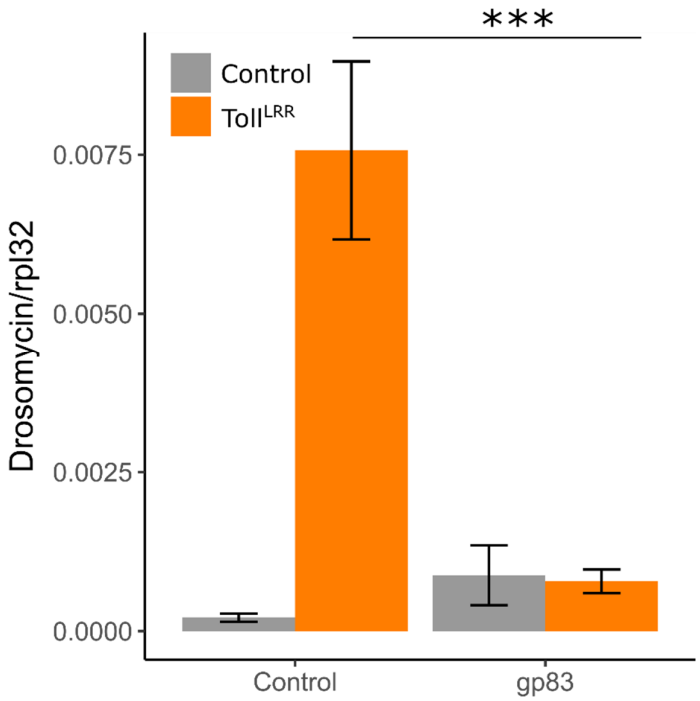

Figure S2: gp83 inhibits endogenous Drosomycin expression

We performed qPCR of Drs, measure relative to rpl32 expression, confirming gp83 is also able to inhibit the expression of endogenously encoded AMPs. Error bars show standard error of the mean (n = 5).

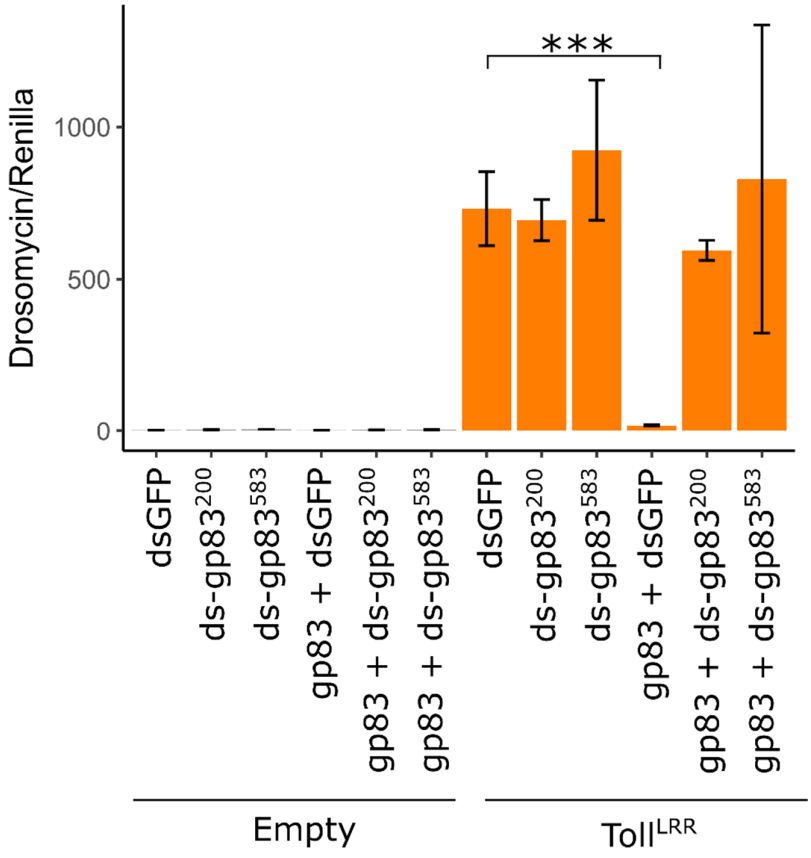

Figure S3: Confirmation of gp83 knock down efficiency

Co-transfection of gp83 with two independent dsRNAs against gp83 was able to reverse the immunosuppressive effect of gp83 on Toll signalling, confirming these dsRNAs were efficient in knocking down gp83 expression. Error bars show standard error of the mean (n = 3).

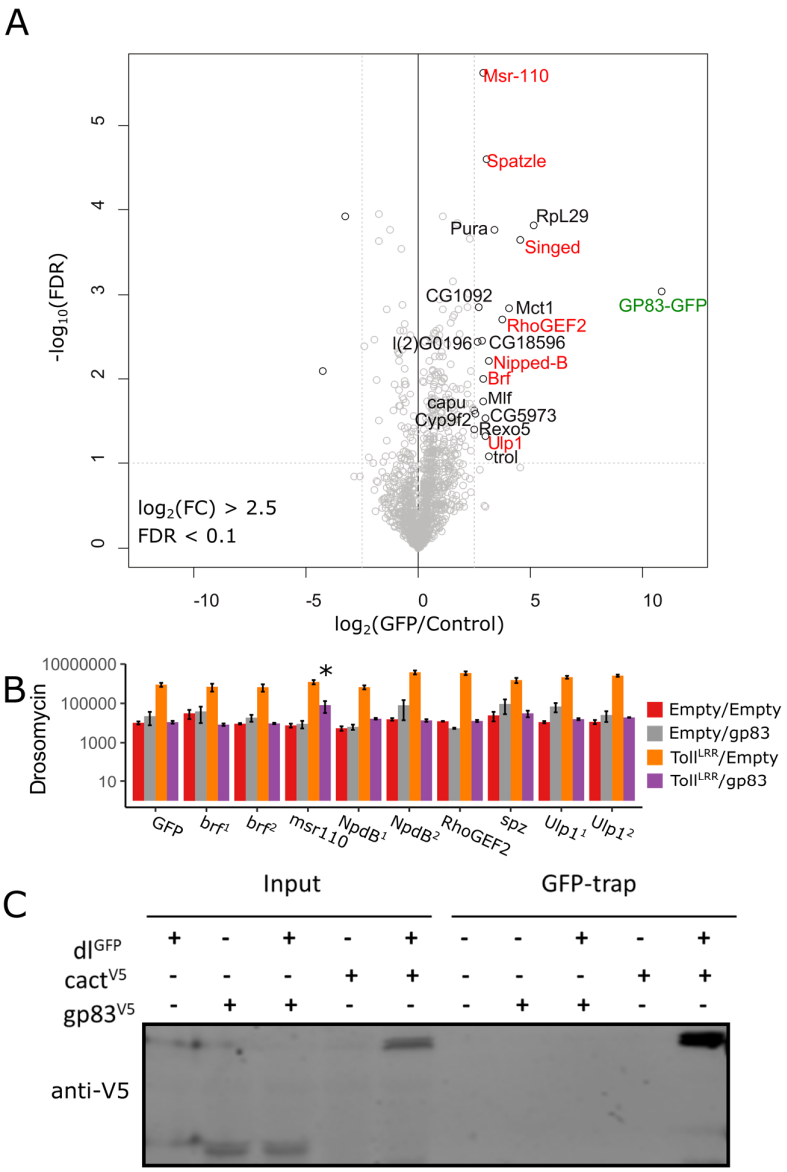

Figure S4: Identification of host interactors of gp83

(A) Identification of gp83 interacting proteins in S2 cell lysates by label-free quantitative (LFQ) mass spectrometry. Permutation-based FDR-corrected t-tests were used to determine proteins that are statistically enriched in gp83-GFP IP. The log2 LFQ intensity of GFP-gp83 IP over control IP (cells that do not express gp83-GFP) is plotted against the -log10 FDR. Interactors with an enrichment of fold change > 2.5; -log10 FDR > 1 are indicated. Although the bait, gp83-GFP (labelled in green), was efficiently retained, few proteins were strongly enriched. (B) Through dsRNA knock-down, we confirmed brf, msr-110, Nipped-B, RhoGEF2, spatzle, and Ulp1 (shown labelled in red in panel A) are not involved in Toll signalling, or gp83 suppression of Toll signalling, as measured by Drs-FLuc expression (relative to pAct-FLuc expression). Genes are superscripted with '1' or '2' when two independent dsRNAs were used to knockdown the gene. Although msr-110 appears to partially rescue gp83 immunosuppression, subsequent experiments did not reproduce this effect. Error bars represent standard error of the mean (n=3). Statistical tests were performed in MCMCglmm. (C) We were unable to identify an association between gp83V5 and dlGFP through dlGFP immunoprecipitation. We used the previously described interaction between cact (V5 tagged) as a positive control. Panel A created by Joep Joosten.

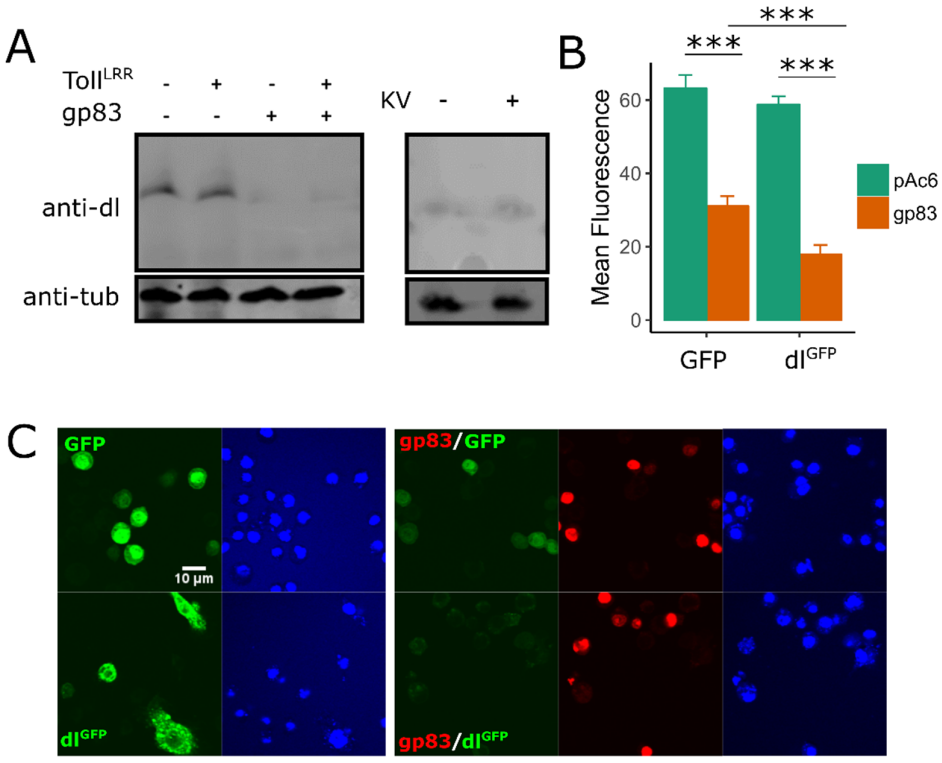

Figure S5: Overexpression of gp83 may reduce dorsal levels

(A) Overexpression of gp83 led to reduced levels of endogenous dl, regardless of Toll activation (i.e. Toll^{LRR} overexpression). (B) However, KV infection did not reduce dl staining, indicating dl degradation may not be the primary mechanism by which KV inhibits Toll signalling (n ≥ 20). (C) Overexpression of gp83 also inhibited both GFP and dl^{GFP} levels, although it reduced dl^{GFP} levels by significantly more (MCMCp < 0.001). Error bars show standard error of the mean (n = 3).