Recurrent evolution of two competing haplotypes in an insect DNA virus

Tom Hill[1]*, Robert L. Unckless[1]

1. 4055 Haworth Hall, The Department of Molecular Biosciences, University of Kansas, 1200 Sunnyside Avenue, Lawrence, KS 66045.

* Corresponding author - Email: tom.hill@ku.edu

Keywords: nudivirus, *Drosophila innubila,* DiNV, virulence, coevolution

Abstract

Hosts and viruses are constantly evolving in response to each other: as hosts attempt to suppress the virus, the virus attempts to evade and suppress the host's immune system. This arms race results in the evolution of novel pathways in both the host and virus to gain the upper hand. Here we describe the coevolution between *Drosophila* species and a common and virulent DNA virus. We identify two distinct viral types that differ 100-fold in viral titer in infected individuals, with similar effects across multiple species. Our analysis suggests that one of the viral types appears to have recurrently evolved at least 4 times in the past ~30,000 years, including in another geographically distinct species, due to the high effective mutation rate which increases with titer. The higher titer viral type is associated with suppression of the host immune system and an increased transmission rate compared to the low viral titer type. Both types are maintained in all populations, likely due to an increased virulence in the high titer type creating a trade-off between effective transmission and virulence and resulting in nearly equal reproduction rates (R_0) in both types. Together these results suggest that the reciprocal selective pressures caused by the co-evolution between host and virus has resulted in this recurrently evolving relationship.

Introduction

Antagonistic coevolution between hosts and their parasites is nearly ubiquitous across the tree of life. As a result, genes involved in immune defense are among the fastest evolving genes in host genomes (NIELSEN *et al.* 2005; SACKTON *et al.* 2007; ENARD *et al.* 2016; SHULTZ AND SACKTON 2019). Viruses are a particular fitness burden on hosts; for viruses to persist within populations, they must successfully invade the host organism, contend with the host immune system, replicate and then transmit the newly produced particles to a new host (HOLMES 2007; GIFFORD 2012). Once successfully established in a population, natural selection acts to modulate the rate the virus propagates relative to its virulence, optimizing the ratio of viral virulence to transmission (WILLIAMS AND NESSE 1991; MAY AND NOWAK 1995; LIPSITCH *et al.* 1996). Thanks to their elevated mutation rate and high population sizes, viruses can easily do all of this, and even co-opt or manipulate host-pathways in the process (BURGYAN AND HAVELDA 2011; DAVEY *et al.* 2011; PALMER *et al.* 2019). As a result, proteins that interact with viruses often show the fastest rates of evolution, and the highest rates of adaptation compared to the rest of the genome, as the host attempts to suppress the pathogen, or escape infection (OBBARD *et al.* 2006; OBBARD *et al.* 2009a; MUKHERJEE *et al.* 2013; ENARD *et al.* 2016; PALMER *et al.* 2018a).

Hosts have evolved numerous immune pathways to reduce viral burden and enhance survival after infection (MERKLING AND VAN RIJ 2013; WEST AND SILVERMAN 2018). Interestingly, several of these pathways (IMD, Toll, JAK-STAT) are generally associated with other pathogens, yet also respond to infection by viruses, though the specific mechanisms are not yet known (HOFFMANN 2003; HULTMARK

2003; COCCIA *et al.* 2004; ZAMBON *et al.* 2005b; COSTA *et al.* 2009; FERREIRA *et al.* 2014a; WEST AND SILVERMAN 2018). In *Drosophila melanogaster*, the RNA interference (RNAi) pathway is involved in resisting viral infection by generating suppressive RNAs complementary to the viral sequence (HULTMARK 2003). As might be expected, this antiviral RNAi pathway is also rapidly evolving in many species (OBBARD *et al.* 2006; WANG *et al.* 2006; OBBARD *et al.* 2009b; MERKLING AND VAN RIJ 2013).

Drosophila innubila Nudivirus (DiNV), which infects the host *D. innubila*, is the one of the few DNA viruses to naturally infect *Drosophila,* and the only documented case of a DNA virus infection at high frequency in natural *Drosophila* populations (UNCKLESS 2011; HILL *et al.* 2019). *Drosophila innubila* is a mushroom feeding species of the *Drosophila* subgenus found inhabiting woodlands found on mountains across Arizona & New Mexico, separated by large expanses of desert. *D. innubila* radiated north from Mexico during the last glaciation period and came to inhabit these forests after the glacial retreat, creating a subdivided population with high rates of gene flow between locations (DYER AND JAENIKE 2005; HILL AND UNCKLESS 2020).

During this period *D. innubila* likely became infected with DiNV, suggesting a long-lasting host-pathogen relationship in multiple populations. This could lead to opportunities to study the coevolution of DiNV and *D. innubila* in replicate (HILL AND UNCKLESS 2020), which could potentially result in parallel or divergent evolution of the virus and interacting host pathways (ANDERSON AND MAY 1982; KALTZ AND SHYKOFF 1998).

A pair of previous studies examined the rates of evolution of DiNV and *D. innubila,* finding the envelope and replication machinery to be rapidly evolving in DiNV, suggesting its importance in viral propagation (HILL AND UNCKLESS 2018). In *D. innubila,* the antiviral RNAi machinery is not rapidly evolving, possibly as DNA viruses interact with different immune pathways to RNA viruses (the primary burden of *D. melanogaster*) (WEBSTER *et al.* 2015). Consistent with this, the Toll pathway is both rapidly evolving in *D. innubila* and suppressed by a related nudivirus upon infection in *D. melanogaster* (HILL *et al.* 2019; PALMER *et al.* 2019).

DNA viruses, and nudiviruses such as DiNV in particular, have much larger genomes (100kbp or larger) than RNA viruses, with much more complicated replication cycles (ROHRMANN 2013). RNA viruses also have much higher mutation rates than DNA viruses, yet much lower levels of diversity due to lower recombination rates and efficient selection on variation in the genome (PENNINGS 2012; PENNINGS *et al.* 2014; WILSON *et al.* 2016; FEDER *et al.* 2019). As a result, we expect the evolutionary dynamics of DNA viruses to differ dramatically from RNA viruses (ROHRMANN 2013; HILL AND UNCKLESS 2017). In fact, because of their large genomes, high recombination rate and low mutation rate, we expect that DNA virus coevolution with hosts will be qualitatively different than RNA viruses or bacteria. Together this paints a picture suggesting that hosts can have different relationships with different pathogens (such as RNA viruses

68 or DNA viruses), and the pathogens themselves can behave differently within the host. Characterizing the
69 relationships between different species and their long-term pathogens, such as DiNV and *D. innubila*, will
70 help broaden and expand our understanding of how hosts and pathogens evolve in response to each other.
71 Here, we survey the genetic variation in DiNV to infer its co-evolutionary history with *D. innubila*
72 and two other associated hosts. We identify two viral multilocus genotypes (considered to be haplotypes)
73 that differ by 11 focal SNPs and show that these viral types are maintained within the same host population
74 and across multiple isolated host populations. Despite high rates of recombination, these SNPs are tightly
75 linked likely due to extremely strong selection and possibly incompatibilities between types. One viral type
76 is associated with 100-fold higher viral titer and increased virulence compared to the other. Further, we find
77 evidence that the high titer type evolved independently in at least four geographically-isolated host
78 populations. Together, these results suggest rapid evolutionary dynamics of host-virus interactions, due to
79 the multiple competing viral types that interact with different host pathways.

80 **Results**

81 *DiNV segregates for linked variants strongly associated with viral titer*

82 To characterize the evolutionary dynamics of wild Drosophila innubila Nudivirus (DiNV) in its host (*D.*
83 *innubila*), we sequenced wild-caught individuals from four populations with the expectation that some
84 (~40% in previous samples) individuals would be infected (HILL AND UNCKLESS 2020). We considered
85 strains to be infected with DiNV if they had at least 10x coverage for 95% of the genome. In total, we used
86 sequencing information for 57, 92, 92, and 92 individuals from the Huachucas (HU), Santa Ritas (SR),
87 Chiricahuas (CH), and Prescott (PR) populations with infection rates 26%, 44%, 63% and 79%, respectively
88 (Supplementary Table 1). We also used 35 individuals collected in the Chiricahuas in 2001 (52% infected
89 with DiNV) and 80 individuals collected in the Chiricahua's in 2018 (Supplementary Table 2, pre-selected
90 using PCR, 40 infected with DiNV and 40 uninfected).
91 We isolated and sequenced DNA from these samples and, after filtering and mapping to the genome
92 we called variation in the viral genomes to assess the extent of adaptation in each viral population.
93 Consistent with an arms race between host and virus, most envelope and novel virulence (GrBNV-like)
94 genes show strong signatures of adaptive evolution in each population compared to background viral genes
95 (using McDonald-Kreitman based statistic Direction of Selection (STOLETZKI AND EYRE-WALKER 2011)
96 and Selection Effect (EILERTSON *et al.* 2012), Supplementary Figure 1, DoS > 0, GLM p-value < 0.05).
97 Given these potential signatures of an arms race between *D. innubila* and DINV, we first attempted
98 to determine if any host or viral genetic variation is associated with within-host viral titer, which we use
99 here as a measure of virulence. For each virus-infected individual, we quantified viral titer (as viral genome
100 coverage normalized to host autosomal genome coverage) and identified both host and viral

101 polymorphisms. We then performed an association study across both host and virus variable sites to identify
102 variants significantly associated with viral titer using PLINK (PURCELL *et al.* 2007).
103 Of 5,283 viral SNPs in the 155kbp DiNV genome, 1,403 SNPs are segregating in at least 5 infected
104 host individuals. Of those 1,403 SNPs, 78 are significantly associated with viral titer after multiple testing
105 correction (FDR < 0.01, Figure 1A). Of these, 16 are within 2000bp of the start site of a gene, 18 are coding
106 nonsynonymous, 11 are coding synonymous and 33 are intergenic. The most significantly associated SNP
107 is a non-synonymous polymorphism in the active site of *Helicase-2* (Figure 1A, Supplementary Figure 2).
108 The *Helicase-2* polymorphism is the only significantly associated polymorphism found segregating at a
109 range of frequencies *within* individuals (Supplementary Figure 2). The frequency of this derived
110 polymorphism has a negative relationship between viral titer and the derived SNPs frequency (GLM t-value
111 = -20.516, *p*-value = 5.55e-62). However, when ranking samples by viral titer, the SNP frequency does not
112 fit this expectation and several samples fixed for the ancestral allele also have a lower viral titer
113 (Supplementary Figure 2).
114 We also identified a striking association between viral titer and eleven nearly perfectly linked
115 polymorphisms found across the DiNV genome (Figure 1A, highlighted SNPs, Table 1, Supplementary
116 Figures 2-4, Sig. SNPs). We refer to these two types as the 'High Type' and 'Low Type' (Figure 1B).
117 This multilocus genotype includes three non-synonymous SNPs and five SNPs in the UTRs of known
118 virulence factor genes and three intergenic SNPs. Viral titer is, on average, 100-fold higher in individuals
119 infected with High type virus compared to the ancestral Low type (Figure 1B). Though we found few
120 strains with an intermediate number of SNPs, viral titer increases as the number of High type SNPs
121 increases (Figure 1C, GLM t-value = 34.971, *p*-value = 5.912e-16), though the rate of increase slows as
122 the number of High type SNPs increases suggesting diminishing returns (Figure 1C). Some of these
123 polymorphisms are associated with known virulence factors, or are related to the formation of the viral
124 envelope co-opting the host vesicle trafficking system and are rapidly evolving in nudiviruses (e.g. *VLF-*
125 *1, ODV-E56, PIF-3*) (ROHRMANN 2013; HILL AND UNCKLESS 2017; HILL AND UNCKLESS 2018).
126 Additionally, several are associated with orphan genes thought to be novel virulence factors, including a
127 *gp83,* a gene that downregulates Toll-induced antimicrobial peptides (AMPs) and upregulates those
128 induced by IMD (*gp83*) (PALMER *et al.* 2019). Both pathways may interact with DNA viruses (ZAMBON
129 *et al.* 2005a; COSTA *et al.* 2009; MERKLING AND VAN RIJ 2013; FERREIRA *et al.* 2014b; LAMIABLE *et al.*
130 2016; PALMER *et al.* 2019).

131 Among populations there is a positive correlation between the frequency of the High type and
132 overall DiNV infection frequency (Figure 1D, GLM logistic regression z-value = 6.104, *p*-value =
133 0.00883), suggesting that the High type may have a higher effective transmission rate, resulting in a

134 higher number of new individuals infected, per DiNV infected individual. The transmission rate appears
135 to be higher for the High type, despite a possibly higher death rate than the Low type. We also find both
136 viral types in collections from 2001 and 2017, with the High type significantly more common in the 2017
137 collection (Fisher Exact Test *p*-value = 0.0167, Figure 1D).

138 **Table 1:** Candidate viral SNPs associated with viral titer, their loci, associated genes and the functional
139 category of that gene.

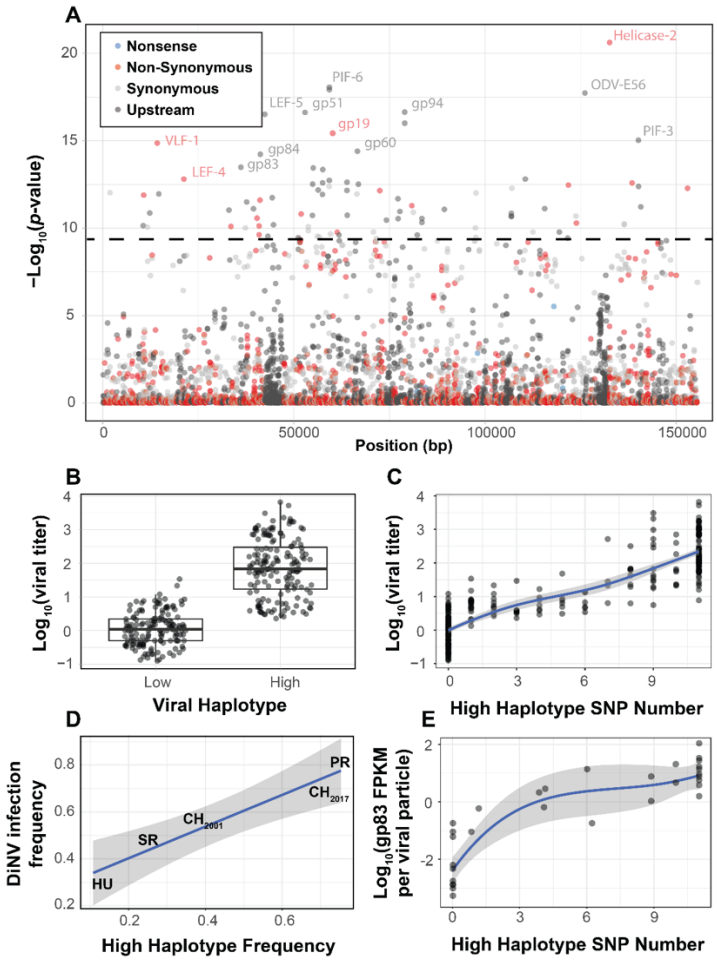
SNP loci	Nearest gene	SNP functional annotation	Nearest gene functional annotation
G14249T	*VLF-1*	Non-Synonymous	Required for late stage virus activity and particle assembly
C41210T	*LEF-4*	Non-Synonymous	RNA polymerase subunit for RNA modification
G42389T	*gp83*	Upstream	Suspected virulence factor which suppresses Toll activity
A59194G	*gp51*	Intergenic	Suspected virulence factor
C59275A	*PIF-6*	Upstream	Per OS Infectivity factor envelope protein, required for oral infection
T59276C	*PIF-6*	Upstream	Per OS Infectivity factor envelope protein, required for oral infection
G66615A	*gp19*	Non-Synonymous	Suspected virulence factor
C78978T	*gp94*	Intergenic	Suspected virulence factor
G78991A	*gp94*	Intergenic	Suspected virulence factor
C126118A	*ODV-E56-2*	Upstream	Occlusion-derived virus envelope protein required for particle formation
A132593C	*Helicase-2*	Non-Synonymous	Unwinds DNA and is critical for DNA replication
T140117C	*PIF-3*	Upstream	Per OS Infectivity factor envelope protein, required for oral infection

140

141 Further, the expression of the virally encoded suppressor of Toll signaling, *gp83,* per viral particle
142 is greater in strains containing at least one derived SNP of the high type increase (Figure 1E, GLM t-value
143 = 10.32, *p*-value = 7.34e-12), suggesting enhanced virulence in the high type. For these SNPs, few are
144 found at intermediate frequencies within samples (Supplementary Figure 2), and no samples contain more
145 than two SNPs at a similar frequency (Supplementary Figure 2). This suggests that hosts are either

146 infected completely with Low type or High type virus particles and the two types may be partially
147 incompatible with each other. Together, these results suggest that the high viral type is more virulent
148 possibly because it is better able to suppress host Toll signaling.

149 **Figure 1:** Viral genome-wide association study for DiNV titer in wild *D. innubila*. **A.** Manhattan plot for
150 each DiNV SNP and the significance of its association with DiNV titer. SNPs are colored if they are
151 upstream, synonymous, non-synonymous or nonsense mutations. 12 named SNPs are either part of the
152 significantly associated viral haplotype or are in *Helicase-2*. The FDR corrected *p*-value cutoff of 0.01 is
153 shown as a dashed line (multiple testing correction for 1,403 tests). **B.** Viral titer for individual wild
154 caught flies infected with Low and High DiNV haplotypes (containing all high-type SNPs). The middle
155 bar represents median value, upper and lower bars represent 25th and 75th percentile and whiskers
156 represent a 95% confidence interval. **C.** Association between the number of significant SNPs and the viral
157 titer of a sample. **D.** Across five populations, the frequency of the High DiNV haplotype is correlated with
158 the frequency of the virus infection. **E.** Expression (in FPKM per viral particle) of *gp83* increases with the
159 number of High DiNV haplotype SNPs.



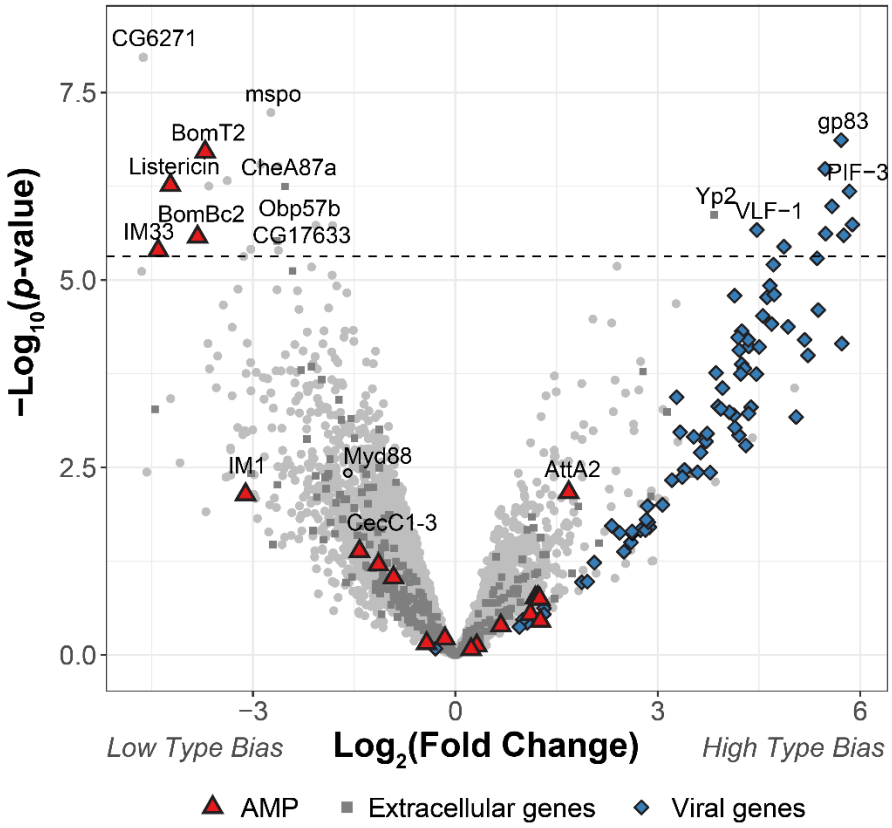
160

161 *The High DiNV viral type suppresses the Drosophila immune system and has elevated virulence*
162 Given the striking difference in viral titer and infection frequencies across populations between High and
163 Low type viruses, we sought to further characterize the differences in infection dynamics between the types.
164 We sequenced mRNA from 80 wild *D. innubila* males collected in 2018 (Supplementary Table 2, 40
165 infected with DiNV, 40 uninfected) and performed a differential expression analysis between infected and
166 uninfected individuals. Few genes were differentially expressed (DE) between infection states in *D.*
167 *innubila* (Supplementary Figure 5A, p-value < 0.01 after FDR multiple testing correction, above the dotted
168 line), but these DE genes were enriched for several interesting categories. Specifically, we found IMD-
169 induced antimicrobial peptides (AMPs) were upregulated upon DiNV infection, while one Toll-induced
170 AMP, and several chorion genes and heat shock proteins were downregulated (Supplementary Figure 5A).
171 We also compared these results to a laboratory experiment in *D. melanogaster* of differential expression
172 after infection with a close relative of DiNV (PALMER *et al.* 2018c). These same genes are also differentially
173 expressed in *D. melanogaster*. Of the 12 genes which are differentially expressed in the same direction in
174 both species, five are AMPs and five are chorion genes (Supplementary Figure 5B).
175 We then compared gene expression between *D. innubila* infected with High or Low type. We find
176 17 DE host genes and 9 DE viral genes between types (after controlling for virus copy number as
177 FPKM/titer, Figure 2, FDR corrected p-value < 0.01, GLM t-value = -4.6239413, p-value = 9.876143e-05).
178 Specifically, four Toll-mediated immune peptides (*Listericin*, *IM33*, Bomanins *BomBC2* and *BomT2*) have
179 reduced expression in High Type infected individuals compared to the Low Type (Figure 2, Supplementary
180 Figure 6). Finally, viral genes of interest (*PIF-3, VLF-1, gp83*) have higher expression per viral particle
181 (FPKM/titer) in the High type compared to the Low type, all also increase in expression per viral particle
182 (FPKM/titer) as the number of High type alleles increases (Figure 1E & 2, t-value = 13.732, p-value 3.36e-
183 15). This pattern may be driven by the allele of the non-synonymous SNP in *VLF-1* (GLM t-value = 2.13,
184 p-value = 0.04272) and the alleles of the SNPs upstream of *gp83*, *PIF-3*, *gp51* and *ODV-E56* (GLM t-value
185 = 3.518, p-value = 0.00162). Together these results suggest that the high viral type has increased expression
186 of key virulence factors, which in turn, manipulate the expression of host genes involved in immune defense
187 to result in the observed differences in viral titer. These results suggest that higher *gp83* expression may
188 cause the lower Toll-mediated AMP expression, possibly due to lowering *Myd88* expression, which in turn
189 prevents the host from enacting a proper immune response to DiNV infection (Figure 2, Supplementary
190 Figure 5).

191 *Experimental infections recapitulate differences in viral type virulence*
192 To assess if the virulence of the virus types differs in experimental infections, we performed experimental
193 infections of *D. innubila* males using viral filtrate of strains infected with one of the two types of DiNV.

194 As infectious viral titer is increased, survival decreases regardless of viral type (Supplementary Figure 7 &
195 8, ANOVA residual deviance = 3.536, *p*-value = 2.454e-07, Cox Hazard Ratio z-value > 2.227, *p*-value <
196 0.02592), with survival decreasing as titer increases (Supplementary Figure 7, Cox Hazard Ratio z-value >
197 5.428, *p*-value < 5.69e-08). In both cases viral titer also increases for the first 3 days of infection (GLM t-
198 value = 9.817, *p*-value = 3.6e-14).
199

200 **Figure 2:** Differential expression of *D. innubila* and DiNV genes between *D. innubila* infected with either
201 the Low type or High type DiNV multilocus genotypes. For host genes, the log-fold change of mRNA
202 fragments per million fragments is compared, while for viral genes the log-fold change of viral mRNA
203 fragments per million fragments per viral particle is compared. Genes are colored/labelled by categories of
204 interest, specifically antimicrobial peptides (AMPs), proteins involved in the extracellular matrix and viral
205 proteins. Specific genes of interest, such as *Myd88*, are also named. The FDR-correct significance cut-off
206 of 0.01 (10,320 tests) is shown as a dashed line.

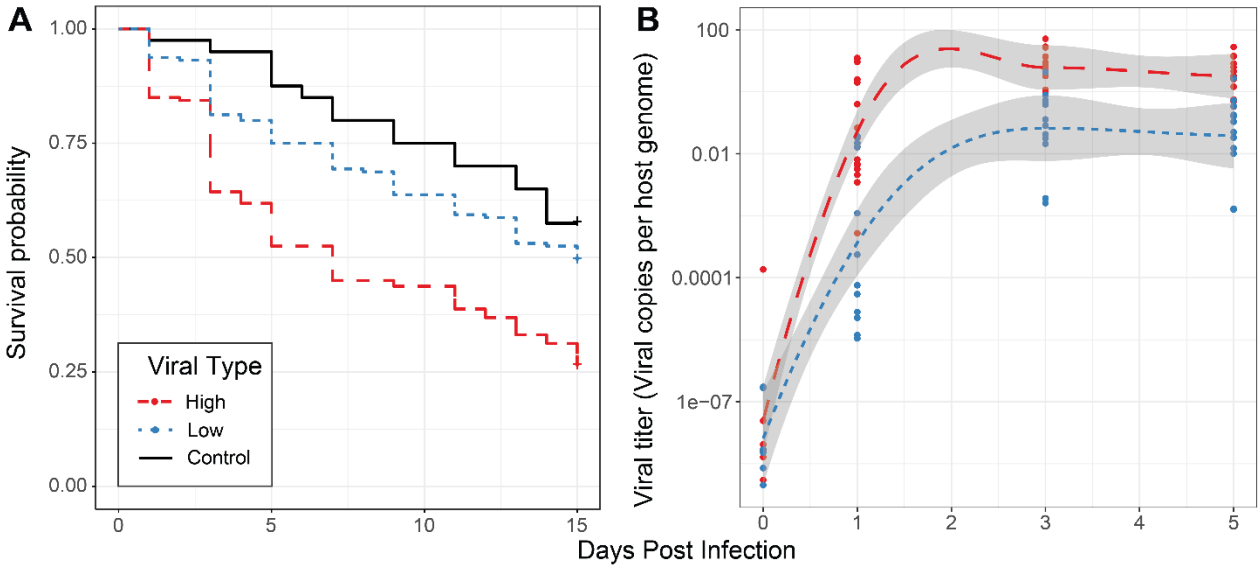


207

208 For a set of independent viral isolates, four High type- and four Low type-infected samples, we
209 diluted samples to roughly equal concentrations of viral particles (about 0.5 particles per host genome prior

210 to filtration) and performed infections for replicates of 10 males with microneedles dipped in one of the
211 filtrate samples. Survival is significantly lower for flies infected with High type viruses when compared to
212 either flies pricked with sterile media (Figure 3A, Cox Hazard Ratio z-value = 3.671 *p*-value = 0.000242)
213 or those pricked with Low type virus (Figure 3A, Cox Hazard Ratio z-value = 4.611 *p*-value = 4e-06). Flies
214 pricked with Low type virus do not show a significant reduction in survival compared to control flies (Cox
215 Hazard Ratio z-value = 1.353, *p*-value = 0.176). We also measured viral titer over time using qPCR, and
216 find titer increases through time in flies infected with either type (Figure 3B, GLM Log_{10}(titer) ~ days +
217 type, days t-value = 9.912, *p*-value = 1.76e-14). Flies infected with High type virus have significantly higher
218 viral titer compared to flies infected with Low type virus (Figure 3B, GLM Log_{10}(titer) ~ days + type, type
219 t-value = 3.934, *p*-value = 0.000211). These results suggest that while the High type strain has higher viral
220 titer and potentially higher transmission rate in wild flies, it also has higher virulence, even after controlling
221 for initial infection titer.

222 **Figure 3:** Effect of viral type in experimental infections. **A.** Survival curves of *D. innubila* infected with
223 high and low viral types compared to control flies pricked with sterile media, for 15 days post infection.
224 Survival 5 days post-infection separated by strain is shown in Supplementary Figure 8. **B.** qPCR copy
225 number of viral *p47* relative to *tpi* in *D. innubila* infected with DiNV filtrate of high and low types.



226

227 *DiNV types are under strong selection in the host*

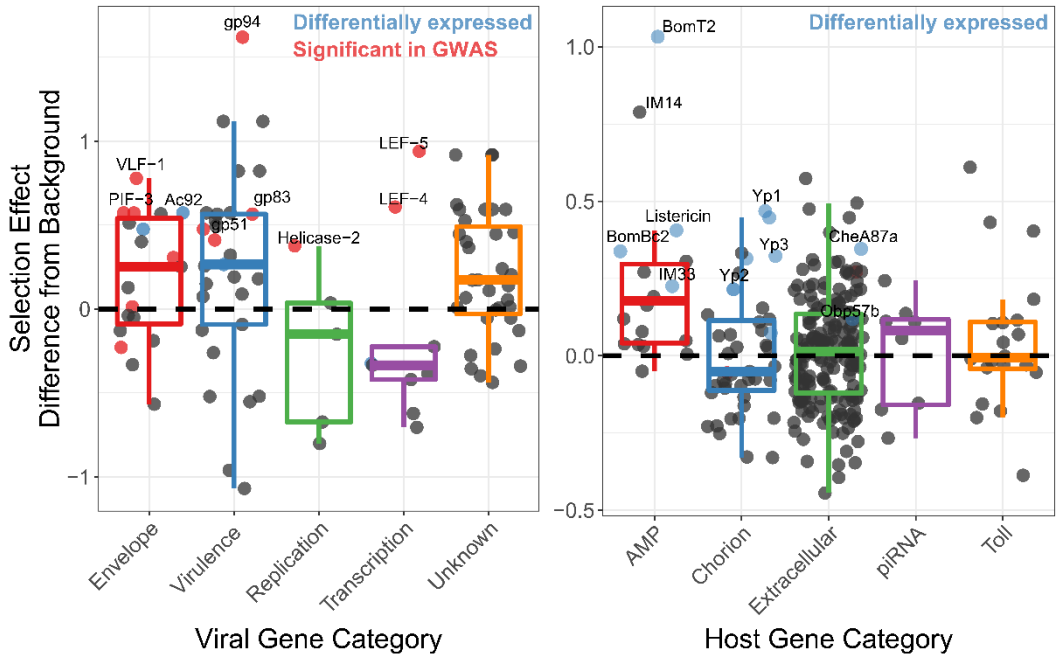
228 Recombination is required during nudivirus replication and recombination start sites can be at any point
229 in the single chromosome circular genome (KELLY 1982; ROHRMANN 2013). These factors likely cause
230 the incredibly high recombination rates observed in nudiviruses (BLISSARD AND ROHRMANN 1990;
231 WANG AND JEHLE 2009; ROHRMANN 2013). In DiNV, the eleven key SNPs that distinguish the High and
232 Low haplotypes are spread across the genome yet are nearly perfectly linked. In contrast, other SNPs in

233 the genome have relatively low linkage disequilibrium, suggesting that selection to maintain each
234 haplotype is strong (Supplementary Figure 4 and 9).
235 Using McDonald-Kreitman based statistics for detecting selection (MCDONALD AND KREITMAN
236 1991; STOLETZKI AND EYRE-WALKER 2011; EILERTSON *et al.* 2012), we tested whether genes that are
237 associated with the High and Low haplotypes exhibited different signatures of natural selection compared
238 to other viral genes. Envelope and virulence proteins show significantly elevated signatures of adaptation
239 (Figure 4, envelope & virulence versus background paired T-test t-value = 2.1761, *p*-value = 0.03814).
240 Genes found in the initial GWAS for virulence, which defined the High and Low types (such as VLF-1,
241 PIF-3, LEF-4 and LEF-5) have significantly higher rate of substitutions being fixed due to selection than
242 background genes (Figure 4¸ type-associated genes versus all other, t-value = 2.718, *p*-value = 0.00068).
243 We also performed a GWAS using the host polymorphism and find few associated SNPs, after
244 controlling for the viral haplotype (Supplementary Figure 10). Consistent with the arms race model, host
245 genes we suspect are interacting with DiNV (such as the GWAS hits, AMPs, chorion genes, piRNA genes
246 and extracellular genes) show elevated levels of substitutions fixed by selection compared to background
247 genes in *D. innubila* (Figure 4 & Supplementary Figure 1). Finally, DE chorion genes, extracellular genes
248 and AMPs have significantly more adaptive substitutions than similar non-DE genes (Figure 4, blue dots,
249 differentially expressed versus all other T-test: *D. innubila* t-value = 4.755, *p*-value = 0.000671). Overall
250 these results suggest strong selection is acting on both the host to suppress viral activity and the virus to
251 escape this suppression.

252 *The High viral type of DiNV evolved repeatedly in three D. innubila populations*
253 We next sought to understand the evolutionary origin of the two types. Given that both types are found in
254 all populations surveyed (Figure 1D), we hypothesized that this could occur one of three ways: First, the
255 derived haplotype was present ancestrally and has been maintained since before geographic isolation
256 occurred. Second, the derived haplotype evolved following geographic isolation and has spread via
257 migration between locations. Third, the derived haplotype has recurrently evolved in each location.
258 To distinguish between these possibilities and determine the timeframe of divergence, we used the
259 site frequency spectrum of silent DiNV polymorphism to estimate effective population size backwards in
260 time for all populations (LIU AND FU 2015). We find three populations (CH, HU and SR) expand from a
261 single viral particle (N_e= 1) to millions of particles during the last glacial maximum (30-100 thousand years
262 ago) when *D. innubila* settled its current range (Supplementary Figure 11). This supports a single invasion
263 event during a host-range change. PR appears to expand between 1 and 10 thousand years ago, suggesting
264 a much more recent bottleneck during the range expansion in PR (Supplementary Figure 11) (HILL AND
265 UNCKLESS 2020).

266 We aligned genomic regions containing SNPs to two related nudiviruses, Kallithea virus and
267 Oryctes rhinoceros Nudivirus (OrNV) (WANG *et al.* 2008; HILL AND UNCKLESS 2018; PALMER *et al.*
268 2018b). The High haplotype alleles are not present in either Kallithea or OrNV, and are not found in short
269 read information for wild *D. melanogaster* infected with Kallithea virus (WEBSTER *et al.* 2015), suggesting
270 they are derived in DiNV.
271

272 **Figure 4:** Genes implicated in host/virus interaction are rapidly evolving by positive selection in the
273 Chiricahua population. Difference in selection effect for viral and host gene categories of interest, and
274 nearby background genes, as indicated by the proportion of substitutions fixed by adaptation, weighted by
275 mutations in SnIPRE (EILERTSON *et al.* 2012). Genes that have associated SNPs from the GWAS are
276 highlighted in red, while genes which are differentially expressed upon infection, or between viral types
277 are labelled in blue. All GWAS hits are also differentially expressed and labelled in red. Genes of interest
278 are named.



279

280 We generated consensus DiNV sequences for each infected *D. innubila* individual and created a
281 whole genome phylogeny to infer geographic diffusion of samples using BEAST (BOUCKAERT *et al.* 2014).
282 We then performed ancestral reconstruction of the presence of the High type across the phylogeny using
283 APE (PARADIS *et al.* 2004). Our samples group as three populations (with HU and SR forming one
284 population) and, consistent with our expectation, the Low type is the ancestral state (Figure 5A).
285 Surprisingly, the High type appears to have evolved repeatedly and convergently within each population,

286 forming separate groups within each population (Figure 5A). The High type also clusters within each
287 population in a principal component analysis of all viral SNPs (Supplementary Figure 9), and when
288 repeating these analyses while excluding the eleven focal SNPs.

289 We next surveyed each background SNP (not associated with the High or Low type) to determine
290 if the general background supports one of the three outlined ways in which the High type evolved and
291 spread in each location. We grouped SNPs by their presence in just the High type or Low type (supporting
292 a single origin and spread by migration) or if they were unique to a single population but shared between
293 the High and Low types (supporting recurrent evolution with recombination).

294 In total, 391 SNPs (28% of SNPs surveyed) are unique to a single population yet are still shared
295 between both High and Low types (Figure 5 and Supplementary Figure 9), compared to 23 SNPs shared
296 between locations but exclusive to High type samples.

297 These unique SNPs (161 for CH, 127 for HU and SR, and 53 for PR), are present in all High type
298 samples of a population but a variable proportion of Low types for that population (between 19% and 94%)
299 and are unique to that population. This pattern fits with the High type recurrently evolving on a single
300 background (a different background in each location), supporting recurrent evolution of the High type. The
301 population-specific background SNPs are spread throughout the DiNV genome, with little evidence of
302 recombination with the high type SNPs, making it unlikely that these SNPs recombined onto different
303 backgrounds (Supplementary Figure 4).

304 Though there is strong linkage between the High type SNPs, they are not perfectly associated with
305 each other (Supplementary Figure 4). Using this slight disassociation and APE (PARADIS *et al.* 2004), we
306 performed ancestral reconstruction of SNP origins in each population (assuming recurrent evolution) and
307 find that, excluding three variable SNPs, the evolution of these SNPs was the same order in each population
308 (Figure 5B).

309 To determine if this recurrent evolution is plausible in our estimated timeframe (~10,000 years),
310 we simulated viral populations using a modified discrete SIR model using deSolve (SOETAERT *et al.* 2010)
311 with estimated baculovirus mutation rates, ranges of viral titer taken our samples and estimated population
312 sizes for each viral population (parameters described in the methods). In this model we used an effective
313 mutation rate scaled to viral titer, considering the mutation rate per particle, so total mutations per
314 generation increase with viral titer. The simulations suggest that waiting time for the first mutation that
315 increases titer is highly variable between replicates but usually occurs within 1000 generations (~200 years
316 assuming 5 generations per year, in >99.1% of replicates, Figure 5C). The average wait time for each
317 subsequent mutation decreases monotonically (GLM t-value = -2.389, *p*-value = 0.03686). In each case,
318 the next mutation appears in the background of the previous high titer mutation (Figure 5C) due to the
319 elevated effective mutation rate and increased basic reproduction number (R_0). The accumulation of

320 mutations therefore occurs at a geometric (approximately exponential) rate. Additionally, the standard
321 deviation of time wait times also decreases with each new mutation (GLM t-value = -2.441, *p*-value =
322 0.04241), increasing the certainty that the entire multilocus genotype will appear in a population rapidly
323 once the initial mutations appear. This chain reaction of adaptation facilitates the repeated evolution of the
324 virulent High type independently in three populations, with all eleven mutations fixing in a population
325 within 6000 generations (~1200 years) in all replicates (3372 generations on average, ~675 years), a
326 plausible amount of time given our estimated timeframe.

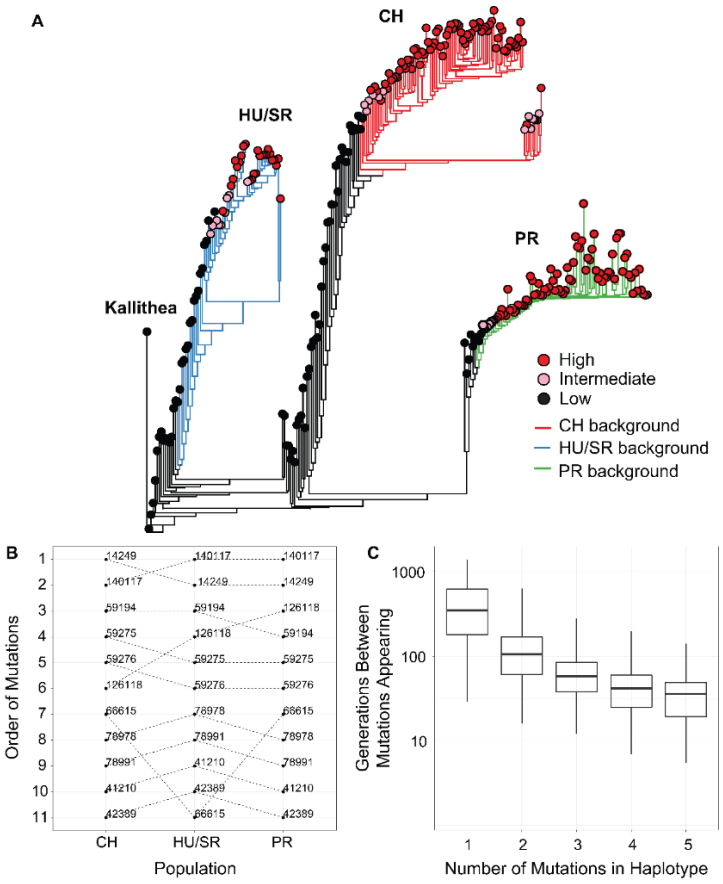
327 *Both viral types are found in two other Drosophila species and have also evolved in a geographically*
328 *distinct population*

329 Since we find two types of DiNV are maintained, and that other species are infected with DiNV, we
330 hypothesized that another species could be a reservoir for the less effective Low type. We chose to study
331 *D. azteca* from the Chiricahuas since it is frequently infected with DiNV (~33% infection), overlaps with
332 *D. innubila,* and is genetically divergent (40-60 million years) which could mean a very different genetic
333 interaction between host and virus (UNCKLESS 2011). We also examined DiNV-infected *D. falleni*
334 (collected in Georgia) as an outgroup. In all, we sequenced 36 *D. azteca* and 56 *D. falleni*. Both types are
335 present in all examined species, but the high type is rare in *D. azteca* (Figure 6B). The High type has a
336 significantly higher titer than the low type in all cases (Figure 6A). Viral titer is not significantly different
337 across species for either High or Low type (Figure 6A, GLM t-value = -1.351, *p*-value = 0.179). We also
338 find the *D. azteca* samples cluster with CH *D. innubila* samples and contain the CH background SNPs
339 (Supplementary Figure 9C), suggesting no differentiation in the virus infecting different species.
340 Interestingly, *D. falleni* DiNV clusters completely separately from the other samples, likely due to its
341 geographic separation, but still has a derived cluster of High type virus, suggesting a fourth separate
342 evolution of the High type in Georgia. Despite the lack of difference between species samples in Arizona,
343 a lower proportion of the *D. azteca* population is infected with DiNV, and the High Type is less common
344 than the Low type DiNV (Figure 6B). Thus perhaps, even though the relative differences in titer are
345 preserved between the two species, the Low Type is favored in *D. azteca* because this reduced virulence
346 leads to a greater R_0 in *D. azteca*. Thus, the two types of the virus may be maintained in both host species
347 because though they have become specialized to maximize fitness in one host, messy transmission between
348 host species could lead to their continued presence in both hosts.

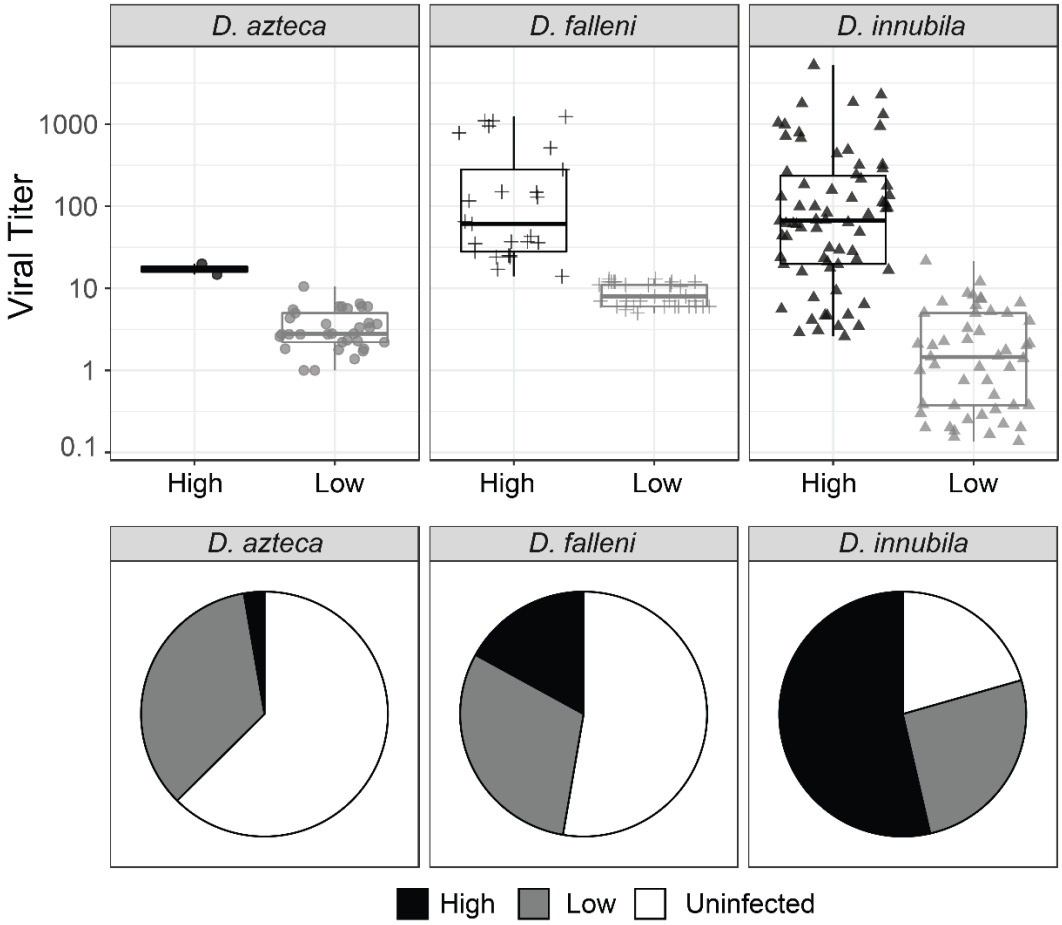
349 We also repeated the GWAS for viral titer in DiNV infecting *D. azteca* and *D. falleni.* In both cases
350 we again found the 11 High type SNPs associated with viral titer (GLM t-value > 4.28, *p*-value > 0.0001 in
351 both cases), but not the *Helicase-2* SNP (despite its presence in *D. falleni* DiNV samples). After controlling
352 for the High type, we find no other significant DiNV SNPs in *D. azteca* associated with viral titer. For

DiNV infecting *D. falleni*, we find 478 significant SNPs (FDR corrected *p*-value < 0.01), though none of them with as large an effect as the High type associated SNPs.

Figure 5: The evolution and maintenance of two viral types. **A.** Phylogeographic reconstruction of the spread of DiNV through *D. innubila,* rooted on the Kallithea virus reference sequence, including a reconstruction of the High type evolution (with strains containing all 11 High type variants shown in red, strains with an intermediate number of high type variants are shown in pink, and strains with no high type variants are shown in black). Branches are colored when the SNPs found in the background for each High haplotype are present in the population, showing that the background differs per population. Black branches show states where branch tips do not contain all the shared High/Low population specific background SNPs. **B.** Order of mutations in the viral haplotype appearing in each population. Apart from three mutations the order is consistent between locations. **C.** The number of generations needed for 'High titer' mutations to evolve in simulated populations, given that each mutation increases the mutation rate. The number of generations between each mutation appearing decreases as titer increases.



369 **Figure 6: A.** Viral titer for CH samples of *D. azteca, D. falleni* and *D. innubila* infected with High and Low
370 type DiNV. **B.** Proportion of *D. azteca* and *D. innubila* 2017 CH population infected with High and Low
371 type DiNV.



372

373 *Two DiNV types may be maintained due to a trade-off between transmission and virulence*

374 Given the ease that the High type appears to evolve recurrently in populations (Figure 5C), its apparent
375 association with increased infection frequency, and its apparent inability to coinfect with the low type, it is
376 surprising that the High type has not outcompeted the Low type. There are several possible explanations
377 for the maintenance of the two types. First, a soft selective sweep may be occurring on the High type, where
378 recurrent mutation followed by a change in environment changes the fitness of the High type that will
379 eventually result in its fixation (HERMISSON AND PENNINGS 2005). Second, both types may be maintained
380 due to a trade-off (ALIZON AND VAN BAALEN 2008). Such a trade-off might even be associated with
381 frequency-dependent selection and cycling frequencies over time. This trade-off could involve different
382 transmission and virulence strategies or might be related to specific adaptation to different hosts.

383 In a simple model of viral infection dynamics, the success of the virus is measured by its basic
384 reproduction number (R_0) which is the ratio of the instantaneous transmission rate (β) to the virulence of
385 the virus (γ). If there is a trade-off between transmission and virulence, we might expect that although the

386 High type has a higher instantaneous rate of transmission, it also has higher virulence, killing infected
387 individuals before they can infect other possible hosts. In contrast, those infected with the Low type persist
388 with the infection and can therefore transmit proportionally more virus due to more interactions with
389 susceptible individuals, despite a lower instantaneous transmission rate. To test this, we simulated
390 populations with two viral types using a modified SIR model in deSolve (SOETAERT *et al.* 2010). We varied
391 transmission and virulence rates and estimated sets of parameters in which types are maintained within
392 populations. We find that a stable infection frequency depends on both the magnitude of instantaneous
393 transmission rate, and the R_0, with higher transmission rates increasing the infection frequency, to a
394 maximum of $1 - (\gamma / \beta)$ (Figure 7A). The total infected proportion effectively saturates due to the increased
395 death rate (virulence) of infected individuals, suggesting a trade-off between transmission and virulence as
396 titer increases (Figure 7A). This is consistent with our results in experimental infections in DiNV
397 (Supplementary Figures 7 & 8, Cox Hazard Ratio z-value > 2.227, *p*-value < 0.02592), and other theoretical
398 treatments (MAY AND NOWAK 1995; ALIZON AND VAN BAALEN 2008). Two types are only maintained
399 when the R_0 is equal for each type ($\gamma_1/\beta_1 = \gamma_2/\beta_2$). As the transmission rate of the High type increases, it
400 infects a larger proportion of the population and outcompetes the Low type (with High and Low types at
401 equal proportions when the transmission rates equal), the High type proportion saturates due to the equally
402 increasing virulence rate which keeps the R_0 equal to the Low type (Figure 7A). Given the requirement for
403 an equal R_0 for maintenance, the actual proportion of individuals infected with each type depends on the
404 starting infection frequencies of each type and the difference in absolute transmission rate (Figure 7B).
405 Based on the infection frequencies of our sampled populations, we suspect that the High type was able to
406 evolve earlier in the recently bottlenecked PR population (consistent with the High type background being
407 shared with 94% of PR Low types), or the absolute transmission rate (and virulence rate) may have
408 increased in PR population, which is likely what has occurred in CH over time (Figure 7B). Together this
409 suggests that differences in population infection frequencies may depend on a combination of demographic
410 factors, host genetic factors and the instantaneous transmission rate in each population (with lower
411 transmission rates in HU and SR compared to PR and CH). This also implies there is a limit to how virulent
412 a strain can become before it becomes detrimental, as even with higher transmission rates per individual,
413 DiNV may kill the host before it can transmit, reducing its basic reproduction number (R_0).

414 **Discussion**

415 Viruses are constantly evolving not just to better infect their host, but also to optimize their infection, to
416 infect as many individuals as possible without preventing the transmission to new hosts (MAY AND NOWAK
417 1995; LIPSITCH *et al.* 1996; ALIZON AND VAN BAALEN 2008). Since the host is also evolving in response
418 to the virus, an evolutionary arms-race often ensues (DAWKINS AND KREBS 1979; KALTZ AND SHYKOFF

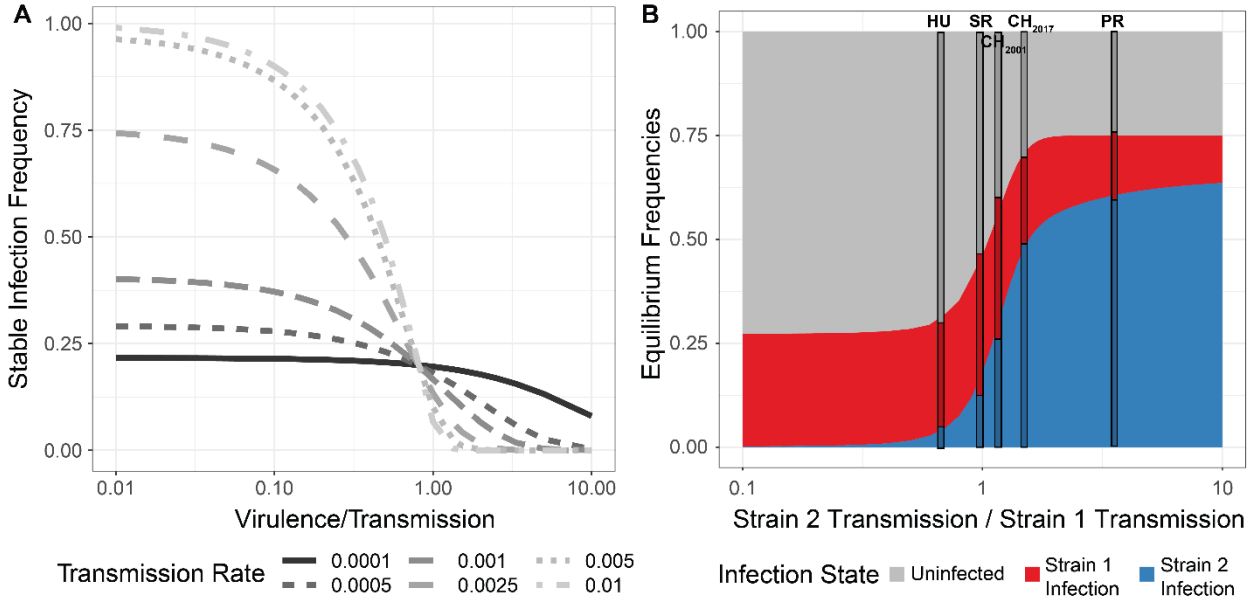
419 1998; DAUGHERTY AND MALIK 2012). Here, to work towards expanding our understanding of the co-
420 evolution of viruses and their hosts, we examine the population dynamics of Drosophila innubila Nudivirus
421 (DiNV), a DNA virus infecting *D. innubila* (UNCKLESS 2011). DNA viruses have large genomes and often
422 recombination, placing them as a somewhat transitionary pathogen between RNA viruses, bacteria and
423 eukaryotic pathogens and parasites. Within our set of viral samples, we find two DiNV haplotypes which
424 differ by 11 SNPs (Figure 1, named High and Low types). One haplotype (the High type) is associated with
425 higher viral titer, likely due to an increased manipulation of the host immune system and increased
426 expression of viral factors. This derived (High type) has likely recurrently evolved in each population since
427 the last glacial maximum (~10,000 years ago). The two types appear to be incompatible to some degree, as
428 we find little evidence of co-infections, and mutations appear in a similar order as if navigating an epistatic
429 fitness landscape (DOBZHANSKY 1937; KONDRASHOV *et al.* 2002; GAVRILETS 2004). Finally, despite the
430 higher titer and transmission rate of the High type, we find that the two types are maintained in all
431 populations sampled, possibly because the increased viral titer also increases the virulence, leading to
432 similar basic reproduction numbers in the High and Low type (ALIZON AND VAN BAALEN 2008).

433 If the reproduction rate, R_0, was equal between the two viral types we would expect the infection
434 frequencies to be equal in our sampled populations, which is not the case. This could be caused by the
435 starting frequencies of each type not being equal (with the Low type starting at a higher frequency, Figure
436 7). There could also be time or host dependent variation, so over time (and between locations), changes in
437 the environment could alter transmission rates of each type, also altering the ratio of High type to Low type.
438 We also do not consider frequency dependent selection in our model, where the transmission rate depends
439 on an interaction between the infection frequencies of the two types.

440 The R_0 could be similar between types, but not identical, resulting in slow changes in the ratio of
441 types over time (as seen between 2001 and 2017, Figures 1D & 7). As we only have two time points, we
442 could be witnessing a selective sweep of the High type spreading to fixation (NIELSEN 2005), with
443 recombination causing the observed differences in the background. As we find the High type appears to
444 have evolved recurrently, it would be unlikely that we have caught these sweeps partway through in all four
445 populations sampled (Figure 1D). Using the frequency of the High type between 2001 and 2017 CH
446 samples, we can calculate the selection coefficient for the High type if increasing at an exponential rate
447 (which is likely given the intermediate frequency at both time points, Figure 7B). If we assume 5
448 generations per year and an increase among infected individuals from 39% on 2001 to 71% in 2017, the
449 selection coefficient = 0.007, which suggests the High type would take ~275 years to fix in a population
450 once it has arisen. Given the coalescence time of the two types is close to the expansion time of the two
451 viruses (2-30 thousand years), this does not fit with our results, suggesting the two types are being

maintained and not sweeping. Further, if the High type was sweeping, it would be remarkable for us to catch these sweeps occurring in all four populations sampled, given our estimated time to fixation.

Figure 7: Simulated infected populations to determine parameters of stable infection frequencies of two viral types. **A.** The frequency of stable infection, given the ratio of virulence to transmission, for different magnitudes of transmission but the same starting frequency (0.001%). **B.** Stable infection frequencies given the difference in transmission rates (β) between two viral strains, as strain 2 transmission rate (starting frequency 0.001%) increases relative to strain 1 transmission rate (starting frequency 25%). The difference in infection frequencies of each strain given differing transmission rates of both strains. Stacked bars show the observed frequencies of High type infection (Strain 2, Blue), Low type infection (Strain 1, Red) and Uninfected (Grey) for four sampled populations, positioned on the X-axis to show the estimated ratio of transmission rates for High and Low type, based on the infection frequencies.



Previous surveys of nudivirus evolution found that a few replication-related genes (including *VLF-1* and *ODV-E56*), likely key targets for host suppression, are under recurrent positive selection (HILL AND UNCKLESS 2017; HILL AND UNCKLESS 2018). Given that these repeatedly rapidly evolving genes are also associated with the High and Low types and show the highest levels of adaptation (Figure 1, Figure 4), it is entirely possible that these genes are key factors for infection in nudiviruses. They may also be associated with virulence in other nudiviruses or baculoviruses. It is interesting that we find the same SNPs recurrently evolving in each population, opposed to different SNPs affecting the same genes in each population.

473 One important factor for DNA virus replication is the *Helicase* gene (BLISSARD AND ROHRMANN
474 1990; ROHRMANN 2013), and we find variation in *Helicase-2* is associated with viral titer in our survey
475 (Figure 3A & C). As in other nudiviruses, there is extensive adaptive evolution in the *Helicase* gene (Figure
476 4A) (HILL AND UNCKLESS 2017; HILL AND UNCKLESS 2018) which has previously been posited to be
477 associated with change in host range (MAEDA *et al.* 1993; CROIZIER *et al.* 1994; ARGAUD *et al.* 1998;
478 AFONSO *et al.* 2001). It is possible the *Helicase-2* variants allow for optimized infection of different species
479 (CROIZIER *et al.* 1994). In fact, DiNV infects several *Drosophila* species in the New World at varying
480 frequencies (including high prevalence in *D. munda*) (UNCKLESS 2011), and so could have an alternate
481 variant reservoir in any species, with some migration into *D. innubila* (but not *D. azteca* or *D. falleni*),
482 resulting in the appearance of two competing strains. However, this hypothesis ignores the existence of the
483 high and low types which appear to explain the difference in viral titer across multiple species much more
484 convincingly than the *Helicase-2* polymorphism (Supplementary Figure 2).
485 Nudiviruses have extremely high rates of recombination, as they require at least one crossover
486 during their replication (ROHRMANN 2013). Given this high rate of recombination, it is interesting we don't
487 find more intermediate strains with a mix of High and Low SNPs, supporting the idea of an incompatibility
488 or negative epistatic interactions between SNPs of the two types, similar to a Dobzhansky-Muller
489 incompatibility (GAVRILETS 2003).
490 Some models suggest that viruses are under constant selection to maintain an optimum ratio of
491 virulence to transmission (MAY AND NOWAK 1995; LIPSITCH *et al.* 1996; ALIZON AND VAN BAALEN 2008).
492 This delicate balance of transmission to virulence could be disrupted with the evolution of a second, high-
493 titer, viral type. In the case of DiNV, the High type has radically higher virulence and appears to compete
494 with the first type impacting the persistence of the virus in a host population. Additionally, the High type
495 appears to recurrently evolve, frequently affecting the persistence of the virus (Figure 5 & 6, Supplementary
496 Figure 9). This posits a situation where the optimum strategy for the ancestral viral type is a reduced
497 transmission rate and the fixation of mutations that are incompatible with High type mutations. Further, this
498 fits with the observed consistent order of fixation for mutations that form the High type (Figure 5B), similar
499 to navigating the neutral adaptive landscape between two incompatible forms in a Bateson-Dobzhansky-
500 Muller incompatibilities (DOBZHANSKY 1937; ORR 1995; ORR AND TURELLI 2001; GAVRILETS 2003; ORR
501 2004). This type of viral type interaction is rarely considered in models used for studying infection and
502 could lead to a better understanding of viral dynamics and host-virus co-evolution (JACKSON *et al.* 2005;
503 JACKSON 2009).
504 DNA viruses such as DiNV have complicated replication cycles and large genomes. This makes
505 them a sort of evolutionary intermediate between RNA-viruses (small genomes, high mutation rates) and
506 eukaryotes (large genomes, low mutation rates) and tangential to bacteria and archaea (intermediate

507 genomes, low recombination rates). However, adaptation appears to occur through changes in a few key
508 proteins. Here we find the evolution of two competing viral types that differ in these few key genes. These
509 viral types are maintained in populations likely due to a trade-off between transmission and virulence.
510 Overall our results suggest that the high mutation rates and extremely high levels of selection can result in
511 the repeated and convergent evolution of novel host-virus interactions. Additionally, we find that these
512 host-virus interactions for large DNA viruses can be much more complicated than previous models suggest
513 (DOLAN *et al.* 2018; FEDER *et al.* 2019).

514 **Materials and Methods**

515 *Fly collection, DNA isolation and sequencing*

516 In this study we used previously collected and sequenced *D. innubila* (HILL AND UNCKLESS 2020). Briefly
517 we collected these flies across the four mountainous locations in Arizona between the 22nd of August and
518 the 11th of September 2017. Specifically, we collected at the Southwest research station in the Chiricahua
519 mountains (~5,400 feet elevation, 31.871 latitude -109.237 longitude), Prescott National Forest (~7,900
520 feet elevation, 34.540 latitude -112.469 longitude), Madera Canyon in the Santa Rita mountains (~4,900
521 feet elevation, 31.729 latitude -110.881 longitude) and Miller Peak in the Huachuca mountains (~5,900 feet
522 elevation, 31.632 latitude -110.340 longitude). Baits consisted of store-bought white button mushrooms
523 (*Agaricus bisporus*) placed in large piles about 30cm in diameter, at least 5 baits per location. A sweep net
524 was used to collected flies over the baits in either the early morning or late afternoon between one and three
525 days after the bait was set. Flies were sorted by sex and species at the University of Arizona and were flash
526 frozen at -80°C before being shipped on dry ice to the University of Kansas in Lawrence, KS. During these
527 collections we also obtained *D. azteca* during collections which we also sorted by species and sex and flash
528 froze. *D. falleni* were collected using a similar method in the Smoky Mountains (~6,600 feet elevation) in
529 Georgia in 2017 by Kelly Dyer, these flies were then sorted at the University of Georgia in Athens GA and
530 shipped on dry ice to the University of Kansas in Lawrence, KS.

531 For collected *D. falleni* and *D. azteca*, we attempted to assess the frequency of DiNV infection
532 using PCR, looking for amplification of the viral gene *p47*. Using primers from (UNCKLESS 2011), P47F:
533 5′–TGAAACCAGAATGACATATATAACGC and P47R: 5′–TCGGTTTCTCAATTAACTTGATAGC.
534 We used the following conditions: 95°C 30 seconds, 55°C 30 seconds, 72°C 60 seconds per cycle for 35
535 cycles.

536 We sorted 343 *D. innubila* flies, 60 DiNV positive *D. falleni* and 40 DiNV positive *D. azteca* which
537 we then homogenized and used to extract DNA using the Qiagen Gentra Puregene Tissue kit (USA Qiagen
538 Inc., Germantown, MD, USA). We prepared a genomic DNA library of these 343 DNA samples using a
539 modified version of the Nextera DNA library prep kit (~ 350bp insert size, Illumina Inc., San Diego, CA,

540 USA) meant to conserve reagents. We sequenced the *D. innubila* libraries on two lanes of an Illumina
541 HiSeq 4000 run (150bp paired-end) (Data to be deposited in the SRA). We sequenced the *D. falleni* and *D.*
542 *azteca* libraries on a separate run of a lane of an Illumina HiSeq 4000 (150bp paired-end).
543 For 80 male *Drosophila innubila* collected in 2018 (indicated in Supplementary Table 2), we split
544 the sample homogenate in half, isolated DNA from half as described above and isolating RNA using the
545 Direct-zol RNA Microprep protocol (R2061, ZymoResearch, Irvine, CA, USA). We then prepared a cDNA
546 library for each of these 80 RNA samples using a modified version of the Nextera TruSeq library prep kit
547 meant to conserve reagents and sequenced these samples on a NovaSeq NS6K SP 100SE (100bp single
548 end). We also sequenced DNA for these samples, with DNA isolated and prepared as above, also sequenced
549 on a NovaSeq NS6K SP 100SE (100bp single end) (Data to be deposited in the SRA).

550 *Sample filtering, mapping and alignment*
551 Following sequencing, we removed primer and adapter sequences using cutadapt (MARTIN 2011) and
552 Scythe (BUFFALO 2018) and trimmed all data using Sickle (-t sanger -q 20 -l 50) (JOSHI AND FASS 2011).
553 We masked the *D. innubila* reference genome (HILL *et al.* 2019), using *D. innubila* TE sequences and
554 RepeatMasker (SMIT AND HUBLEY 2008; SMIT AND HUBLEY 2013-2015). We then mapped short reads to
555 the masked genome and the Drosophila innubila Nudivirus genome (DiNV) (HILL AND UNCKLESS 2018)
556 using BWA MEM (LI AND DURBIN 2009) and sorted using SAMtools (LI *et al.* 2009). Following this we
557 added read groups, marked and removed sequencing and optical duplicates, and realigned around indels
558 in each mapped BAM file using GATK and Picard (HTTP://BROADINSTITUTE.GITHUB.IO/PICARD ;
559 MCKENNA *et al.* 2010; DEPRISTO *et al.* 2011). We considered lines to be infected with DiNV if at least
560 95% of the viral genome is covered to at least 10-fold coverage. We then filtered for low coverage and
561 mis-identified species by removing individuals with low coverage of the *D. innubila* genome (less than 5-
562 fold coverage for 80% of the non-repetitive genome), and individuals we suspected of being misidentified
563 as *D. innubila* following collection. This left us with 318 *D. innubila* wild flies with at least 5-fold
564 coverage across at least 80% of the euchromatic genome, of which 254 are infected with DiNV
565 (Supplementary Table 1). We also checked for read pairs which were split mapped between the DiNV
566 genome and the *D. innubila* genome using SAMtools.
567 For *D. falleni* we used a previously generated *D. innubila* genome with *D. falleni* variants
568 inserted (HILL *et al.* 2019). We masked the genome with Repeatmasker (SMIT AND HUBLEY 2013-2015)
569 and mapped short reads to the masked genome, the repeat sequences and the DiNV genome using BWA
570 MEM and SAMtools (LI AND DURBIN 2009; LI *et al.* 2009). Then, as with *D. innubila* we filtered for low
571 coverage and mis-identified species by removing individuals with low coverage (less than 5-fold

572 coverage for 80% of the non-repetitive genome) leaving us with 56 *D. falleni* samples infected with
573 DiNV.
574 For *D. azteca*, we downloaded the genome from NCBI (Accession: GCA_005876895.1) which
575 we then called repeats from with RepeatModeler (SMIT AND HUBLEY 2008). We masked the genome with
576 Repeatmasker (SMIT AND HUBLEY 2013-2015) and mapped short reads to the masked genome, the repeat
577 sequences and the DiNV genome using BWA MEM and SAMtools (LI AND DURBIN 2009; LI *et al.*
578 2009). As with *D. innubila* we then filtered for low coverage and mis-identified species by removing
579 individuals with low coverage of the *D. azteca* genome (less than 5-fold coverage for 80% of the non-
580 repetitive genome), which left us with 37 *D. azteca* samples infected with DiNV. We then called DiNV
581 variation using LoFreq (WILM *et al.* 2012).

582 *Calling nucleotide polymorphisms across the population samples*
583 For the 318 sequenced samples with reasonable coverage, for host polymorphism, we used the previously
584 generated multiple strain VCF file, generated using a standard GATK HaplotypeCaller/BCFTools pipeline.
585 We used LoFreq (WILM *et al.* 2012) to call polymorphic viral SNPs within each of the 254 DiNV infected
586 samples, following filtering using BCFtools to remove sites below a quality of 950 and a frequency less
587 than 5%. We then merged each VCF to create a multiple strain VCF file, containing 5,283 SNPs in the
588 DiNV genome. The LoFreq VCF (WILM *et al.* 2012) output contains estimates of the frequency of each
589 SNP in DiNV in each sample, to confirm these frequencies, in SAMtools (LI *et al.* 2009) we generated
590 mPileups for each sample and for SNPs of interest (related to viral titer), we counted the number of each
591 nucleotide to confirm the estimated frequencies of these nucleotides at each position in each sample. To
592 confirm that there are no coinfections of types, we also subsampled samples and randomly merged low and
593 high type samples and again generated mPileup files, for SNPs of interest we again counted the number of
594 each nucleotide at each position and confirmed these matched our expected counts in the merged files. We
595 then compared these artificial coinfections to actual samples to confirm the presence or absence of
596 coinfections, finding no samples consistent with coinfections. We then used SNPeff to identify the
597 annotation of each SNP and label synonymous and non-synonymous (CINGOLANI *et al.* 2012). We extracted
598 the synonymous site frequency spectrum to estimate the effective population size backwards in time using
599 StairwayPlot (LIU AND FU 2015).

600 *Identifying differentially expressed genes between DiNV infected and uninfected Drosophila innubila*
601 For 100 male *Drosophila innubila* collected in 2018 (indicated in Supplementary Table 2), we
602 homogenized each fly separately in 100µL of PBS. We then split the sample homogenate in half, isolated
603 DNA from half as described above and isolating RNA using the Direct-zol RNA Microprep protocol
604 (R2061). Using the isolated DNA, we tested each sample for DiNV using PCR for *P47* as described

605 previously, using 40 DiNV infected samples and 40 uninfected samples. We then prepared a cDNA
606 library for each of these 80 RNA samples using a modified version of the Nextera TruSeq library prep kit
607 meant to conserve reagents and sequenced these samples on a NovaSeq NS6K SP 100SE (100bp single
608 end). We also sequenced DNA for these samples, with DNA isolated and prepared as above, also
609 sequenced on a NovaSeq NS6K SP 100SE (100bp single end). The DNA sequenced here was mapped as
610 described above, with variation called as described above for other DNA samples.
611 Following trimming and filtering the data as described in the methods, we mapped all mRNA
612 sequencing data to a database of rRNA (QUAST *et al.* 2013) to remove rRNA contaminants. Then we
613 mapped the short read data to the masked *D. innubila* genome and DiNV genome using GSNAP (-N 1 -o
614 sam) (WU AND NACU 2010). We estimated counts of reads uniquely mapped to *D. innubila* or DiNV
615 genes using HTSEQ (ANDERS *et al.* 2015) for each sample. Using EdgeR (ROBINSON *et al.* 2009) we
616 calculated the counts per million (CPM) of each gene in each sample and counted the number of samples
617 with CPM > 1 for each gene. We find that over 70.3% of genes have a CPM > 1 in at least 70 samples.
618 For the remaining genes, we find these genes are expressed in all samples of a subset of the strains (e.g.
619 DiNV uninfected, DiNV infected, DiNV high infected, DiNV low infected). This supports the validity of
620 the annotation of *D. innubila*, given most genes are expressed in some manner, and suggests our RNA
621 sequencing samples show expression results consistent with the original annotation of the *D. innubila*
622 genome.
623 We attempted to improve the annotation of the *D. innubila* genome to find genes expressed only
624 under infection. We extracted reads that mapped to unannotated portions of the genome and combined these
625 for uninfected samples, samples infected with high type DiNV and samples infected with low type DiNV
626 as three separate samples. We then generated a *de novo* assembly for each of these three groups using
627 Trinity and Velvet (SCHULZ *et al.* 2012; HAAS *et al.* 2013). We then remapped these assemblies to the
628 genome to identify other transcripts and found the consensus of these two for each sample. Using the
629 Cufflinks pipeline (GHOSH AND CHAN 2016), we mapped reads to the *D. innubila* genome and counted the
630 number of reads mapping to each of these putative novel transcript regions, identifying 15,676 regions of
631 at least 100bp, with at least 1 read mapping in at least 1 sample. Of these, 717 putative genic regions have
632 at least 1 CPM in all 80 samples, or in all samples of one group (DiNV uninfected, DiNV infected, DiNV-
633 low infected, DiNV-high infected). We next attempted to identify if any of these genes are differentially
634 expressed between types, specifically between uninfected strains and DiNV infected strains, and between
635 low-type infected and high-type infected strains. Using a matrix of CPM for each putative transcript region
636 in each sample, we calculated the extent of differential expression between each type using EdgeR
637 (ROBINSON *et al.* 2009), after removing regions that are under expressed, normalizing data and estimating
638 the dispersion of expression. We find that 26 putative genes are differentially expressed between infected

639 and uninfected types, and 69 putative genes are differentially expressed between high and low types. We
640 took these regions and identified any homology to *D. virilis* transcripts using blastn (ALTSCHUL *et al.* 1990).
641 We find annotations for 37 putative genes are either expressed in all samples, or differentially expressed
642 between samples. Of the 14 putative genes expressed in all samples, nine have the closest blast hit to an
643 rRNA gene, and five have hits to unknown genes. For 23 differentially expressed putative genes with blast
644 hits, 3 genes are like antimicrobial peptides (*IM1, IM14, IM3*), these genes are significantly downregulated
645 upon infection, like other Toll regulated AMPs, and have significantly lower expression in strains infected
646 with high type DiNV compared to low types. The remaining 20 genes all have similarity to genes associated
647 with cell cycle regulation, actin regulation and tumor suppression genes.

648 *Identifying genes associated with viral titer in Drosophila innubila*

649 As the logarithm of viral titer was normally distributed (Shapiro-Wilk test W = 0.05413, *p*-value = 0.342),
650 we used PLINK (PURCELL *et al.* 2007) to associate nucleotide polymorphism to logarithm of viral titer in
651 infected samples. We fit a linear model in PLINK including population, sex, *Wolbachia* presence, the date
652 of collection and the relationship matrix for relationship of each sample (inferred using PLINK).

653 We first fit this model for all 5,283 viral polymorphisms, before performing the association study,
654 we also pruned viral SNPs for both the total population and each subpopulation leaving 1,403 SNPs. For
655 the total sample we identified associations between the logarithm of viral titer and the frequency of the viral
656 polymorphism in each individual sample, resulting in the following model:

657 $$Log_{10}(viral\ titre) \sim SNP + hs + w + p + dc + (SNP * hs) + (SNP * p) + (SNP * w)$$
658 $$+ relationship[strain]$$

659 Where hs = host sex, p = location of collection, w = *Wolbachia* presence, dc = date collected

660 Following model fitting, we found factors which seemed to show little or no effect on viral titre (*p*-value >
661 0.1) using an ANOVA in R (TEAM 2013), and removed these, refitting the model. This was done step-wise,
662 leaving the following model by the end:

663 $$Log_{10}(viral\ titre) \sim SNP + hs + p + (SNP * hs) + relationship[strain]$$

664 Following this we also performed a GWAS using PLINK (PURCELL *et al.* 2007) in the host, using
665 previously called host variation, and considering viral haplotype as an additional covariate.

666 $$Log_{10}(viral\ titre) \sim SNP + hs + p + (SNP * hs) + vh + relationship[strain]$$

667 Where hs = host sex, p = location of collection, w = *Wolbachia* presence, dc = date collected, vh = viral
668 haplotype. We found no convincing significant associations (Supplementary Figure 11).

669 We repeated this analysis for DiNV variants in *D. azteca* and *D. falleni* separately. We performed
670 the GWAS twice, first using the original model, then including viral haplotype as an additional covariate.

671 *Estimating viral titre using qPCR*

672 Following the identification of the viral haplotype associated with viral titre we sought to determine the
673 effect of viral haplotypes in actual infections. For 20 samples with fly homogenate, we determine the viral
674 titer and haplotype following filtration with a 0.22µM filter.

675 We performed qPCR for the viral gene *p47* (Forward 5-TCGTGCCGCTAAGCATATAG-3,
676 Reverse 5-AAAGCTACATCTGTGCGAGG-3) on 1 µL of fly filtrate per sample and compared the
677 estimated Cq values across 3 replicates to estimated viral copy number to confirm viral concentration
678 (protocol: 2 minutes at 95°C, 40 cycles of 95°C for 30 seconds and 59°C for 20 seconds, followed by 2
679 minutes at 72°C). Following this we diluted samples to similar Cq values, relative to the sample with the
680 highest Cq value. We confirmed this by repeating qPCR with *p47* primers of 1 µL of each sample.

681 For each filtrate sample we performed infections on 30 *D. innubila* males 4-5 days following
682 emergence using pricks with sterile needles dipped in viral filtrate. We recorded survival of each fly each
683 day and removed dead flies. Finally, we took samples 1, 3 and 5-days post infection and measured viral
684 copies of *p47* relative to *tpi* at each time point.

685 *Phylogeography of DiNV infection and the evolution of the different viral haplotypes*

686 For each DiNV infected *D. innubila* sample, we reconstructed the consensus DiNV genome infecting them
687 using GATK AlternateReferenceMaker and the VCF generated for each strain (McKenna *et al.* 2010;
688 DePristo *et al.* 2011). We then converted these genomes into Phylip format and used BEAST2 to build
689 the phylogeny of DiNV genomes using 100 million iterations with a burn in of 5 million (Bouckaert *et*
690 *al.* 2014). We considered phylogeography by providing the longitude and latitude of each samples
691 collection. We then generated a final consensus phylogeny using Tracer with at least 90% majority
692 consensus and removed trees that had not converged to the same joint density (Bouckaert *et al.* 2014).
693 To reconstruct the evolution of the high type viral type we used APE (Paradis *et al.* 2004) to infer the
694 appearance order of the six perfectly linked SNPs on the phylogeny using the all different rates (ARD)
695 discrete model (Paradis *et al.* 2004). We also confirmed these recurrent mutations across the phylogeny
696 using TreeTime to identify recurrently evolving SNPs (Sagulenko *et al.* 2018). Finally, we also created
697 a matrix of SNPs present in at least 2 viral samples and used this matrix in a principle component analysis
698 in R (Team 2013), labelling each sample by their viral type in the PCA.

699 *Simulating the evolution of the high and low viral haplotypes.*

700 We sought to simulate the infection of DiNV in *D. innubila* when considering the evolution of a high titer
701 viral haplotype, specifically if two viral types can be maintained against each other at stable frequencies,
702 and if the high viral haplotype with 5 shared mutations could evolve recurrently in the given time period
703 given realistic parameters. We used the R package DeSolve (Soetaert *et al.* 2010) to simulate infection

704 dynamics in a modified SIR model. Specifically, we removed a resistant class, under the assumption that
705 flies won't live long enough to shed the infection. Therefore, the proportion of population infected per
706 generation is as follows:

707 $$p_{Infected} = (Susceptible * Infected * \beta) - (Infected * \gamma)$$

708 Where β represents an infection parameter and γ represents a virulence parameter (e.g. the increased
709 likelihood an infected individual has of dying before it can spread its infection). This equation can be
710 rearranged to show the stabilized maximum frequency of infection:

711 $$p_{Infected} = 1 - \frac{\gamma}{\beta}$$

712 Which is maintained while γ is greater than 0.01. The average frequency decreases as absolute transmission
713 rate decreases past this point. We then extended this to include two competing infection types, to represent
714 the two viral types, with the total proportion of population infected per generation as follows:

715 $$p_{Susceptible} = 1 - (Susceptible * Infected_1 * \beta_1) - (Infected_1 * \gamma_1)$$
716 $$+ (Susceptible * Infected_2 * \beta_2) - (Infected_2 * \gamma_2)$$

717 This equation can be rearranged as before to show the stabilized frequency of infection for two viral types:

718 $$p_{Infected} = 1 - \frac{\gamma_1 + \gamma_2}{\beta_1 + \beta_2}$$

719 In this case, the two viral haplotypes can both be maintained within a population when the ratio of infection
720 to virulence are the same:

721 $$\frac{\gamma_1}{\beta_1} = \frac{\gamma_2}{\beta_2}$$

722 We assessed how starting frequency and difference in γ and β affects the evolution of each type and their
723 stable frequencies but repeating these simulations for 10,000 generations for each set of parameters. We
724 repeated simulations with transmission rates varying between 0.00001 and 1, and virulence rates set as the
725 transmission rate, divided by a scaling factor between 0.01 and 10 (to vary virulence at multiple rates higher
726 and lower than transmission rate), with all pairwise combinations across all parameters.

727 Next, we attempted to assess if the 'high titer' viral haplotype could possibly evolve recurrently in
728 each population in the time scale seen in our findings. We again used the modified SIR model, this time
729 discrete with population sizes set to 1 million individuals, based on *StairwayPlot* estimates (LIU AND FU
730 2015). For each infected individual in the population, we tracked the viral titer and also recorded the
731 presence of absence of five mutations, with each mutation increasing the titer of infection, but each further
732 mutation having successively smaller increases in viral titer ($titer^{\sqrt{no.muts}}$), representing the epistatic
733 interaction of high titer associated mutations seen in the viral haplotype. We multiplied the infection
734 parameter, mutation parameter and virulence parameter by viral titer, under the assumption that viral titer
735 increases both infection and death rate, and the mutation rate is per viral particle. We considered a per site

736 mutation rate of 10^{-6}, based on estimated baculovirus mutation rate (ROHRMANN 2013; CHATEIGNER *et al.*
737 2015), with a specific mutation rate for the five haplotype mutations of 6.4e-12 (1e-6/155kbp) * viral titer.
738 We then simulated populations in replicate 1000 times for 100,000 generations with a starting infection
739 frequency of 10% for the 'low titer' haplotype, recording the frequency of the virus in a population, the
740 frequency of the haplotype and the time that each 'high titer' mutation reaches high enough frequency to
741 escape stochastic behavior and behave deterministically under selection (GILLESPIE 2004).
742 To estimate the possible selection coefficient for DiNV in the CH population, we assumed an
743 exponential distribution and 5 viral generations per year (80 generations between 2001 and 2017). We then
744 solved the following equation:

745 $$P_{2017} = P_{2001} * (1 + s)^t$$

746 Where P_{2017} = the frequency of the High type among viral samples in 2017 (71%), P_{2001} = the frequency of
747 the High type among viral samples in 2001 (39%), s = the selection coefficient and t = the number of
748 generations (80). We then used this estimated selection coefficient in the same equation to find the number
749 of generations (t) to go from $1/2N_es$ (0.0000714, assuming an N_e of 1000000) to fixation (0.99):

750 $$0.99 = 0.0000714 * (1 + 0.007)^t$$

751 *Experimental infections of Drosophila innubila with DiNV*
752 We chose *D. innubila* samples infected with DiNV and with sequenced genomes, 4 infected with the high
753 type DiNV and 4 infected with the low type. For these samples we estimated their viral copy number per
754 host genome as described previously. We used qPCR on *p47* and *tpi* to find the differences in Cq values to
755 calculate the concentration of each sample relative to the lowest concentration sample and diluted 50µL of
756 filtrate for each sample to match the concentration of each sample to the samples with the lowest titer
757 (IPR07). For a separate 50µL of the IPR01 sample, we performed 1 in 10 serial dilutions to give 45µL of
758 filtrate at full concentration, 1 in 10 concentration, 1 in 100 concentration and 1 in 1000 concentration.
759 Using these sets of samples (matched titer and serial dilutions) we next performed experimental infections.
760 We transferred 50 *D. innubila* (of roughly equal sex ratio) to new food and let them lay eggs for 1
761 week, following this we collected male offspring aged 2-5 days for experimental infections.
762 Across 4 separate days in the mornings (between 9am and 11am), we infected the collected male
763 flies with each sample. For flies in batches of 10, we performed pricks with microneedles dipped in the
764 prepared viral filtrate. For each day we also had 2 control replicates of 10 flies pricked with microneedles
765 dipped in sterile media. Following infections, we checked on each vial of 10 flies one-hour post infection
766 and removed dead flies (likely killed by the needle instead of the virus). We also checked each vial each
767 morning for 15 days, removing dead flies (freezing to determine the viral titer), and flipping flies to new

768 food every 3-4 days. Checking at 10am each day, we recorded the day that each fly died, what filtrate they
769 had been infected with, and what replicate/infection day set they belonged to. We next looked for
770 differences in survival over time compared to sterile wound controls using a Fit proportional hazards
771 regression model in R (TEAM 2013; KASSAMBARA *et al.* 2017), considering titer, infection sample and
772 replicate as co-variates (day of death ~ [titer or strain] + infection date).

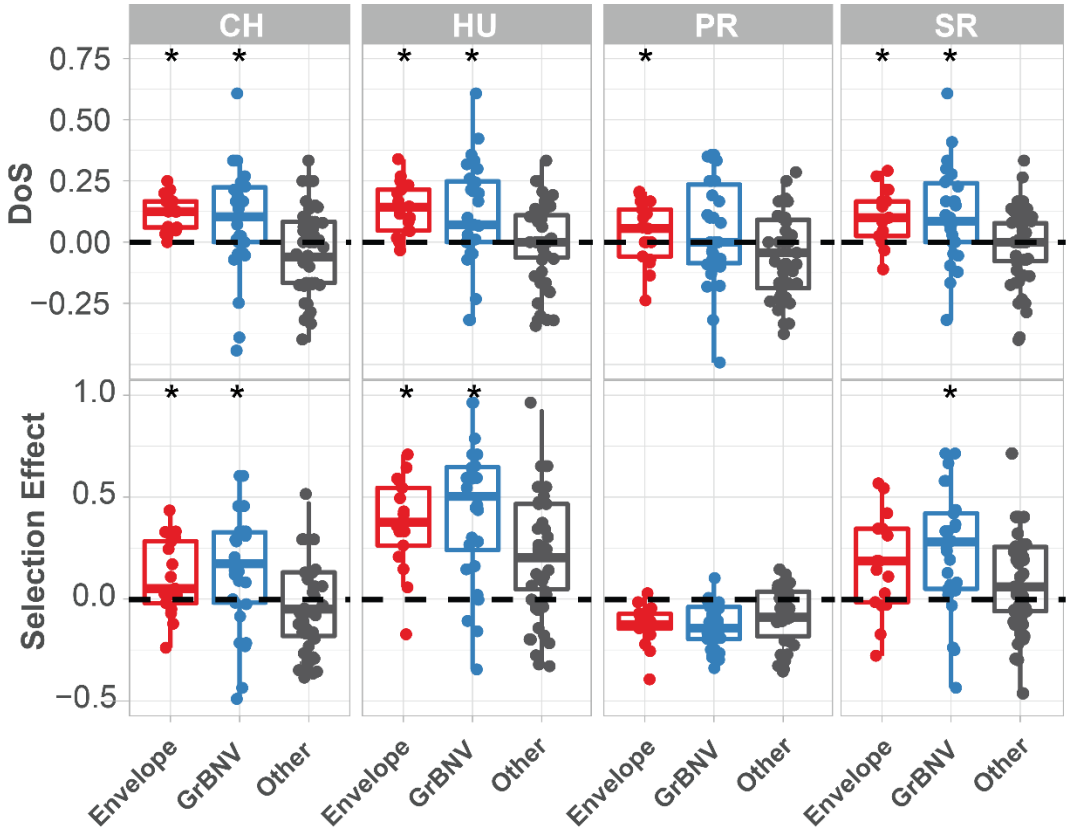
773 For a second set of experimental infections (performed as described above, stabbed with diluted
774 filtrate from different strains), we also removed 3 living flies 1 hour, 1 day and 5 days post infection. Using
775 qPCR, we found the difference in *p47* log-Cq and *tpi* log-Cq to estimate the viral copy number for each
776 sample over time.

777 **Acknowledgements**

778 This work was completed with helpful discussion from Justin Blumenstiel, Joanne Chapman, John Kelly,
779 Stuart MacDonald, Andrew Mongue and Carolyn Wessinger. We would especially like to thank Maria
780 Orive, Kelly Dyer and Paul Ginsberg for helpful feedback in the writing of the manuscript and framing of
781 the discussion. Collections were completed with assistance from Todd Schlenke, Paul Ginsberg, Kelly Dyer,
782 Brandon Cooper, John Jaenike and the Southwest Research Station. We thank Brittny Smith and Jenny
783 Hackett at the KU CMADP Genome Sequencing Core (NIH Grant P20 GM103638) and K-INBRE
784 Bioinformatics Core for assistance in genome isolation, library preparation, sequencing and computational
785 resources. This work was supported by a K-INBRE postdoctoral grant to TH (NIH Grant P20 GM103418).
786 This work was also funded by NIH Grants R00 GM114714 and R01 AI139154 to RLU. *D. falleni* collection
787 was funded by NSF grant DEB-1737824.

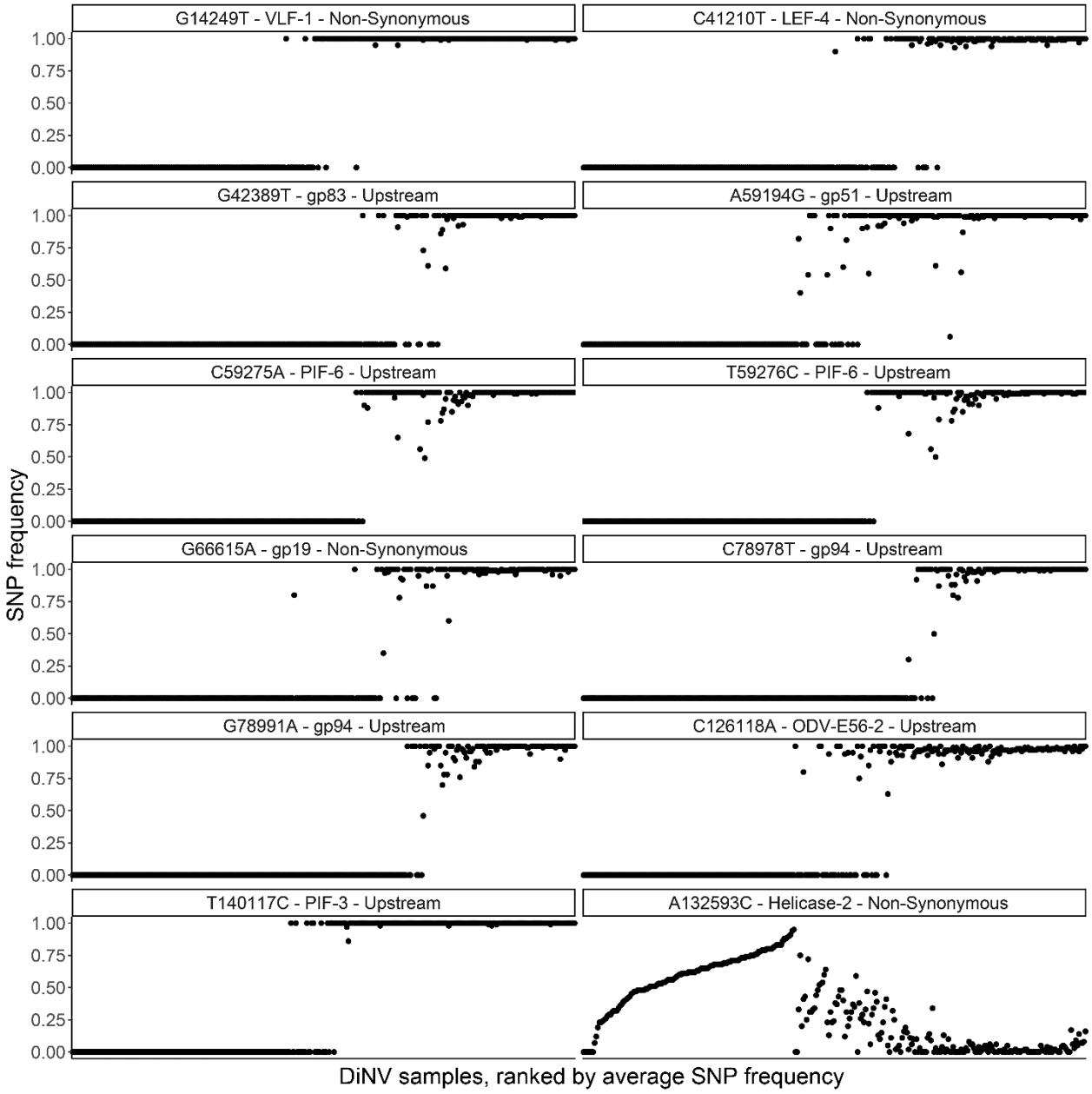
788 **Supplementary Figures**

789 **Supplementary Figure 1:** McDonald-Kreitman based statistics for each gene in population of Drosophila
790 innubila Nudivirus, with viral envelope and GrBNV potential virulence factors shown separately. DoS =
791 direction of selection, Selection Effect = SnIPRE estimated weighted DoS. Boxplots marked with a * are
792 significantly higher than background/other viral genes (GLM *p*-value < 0.05).



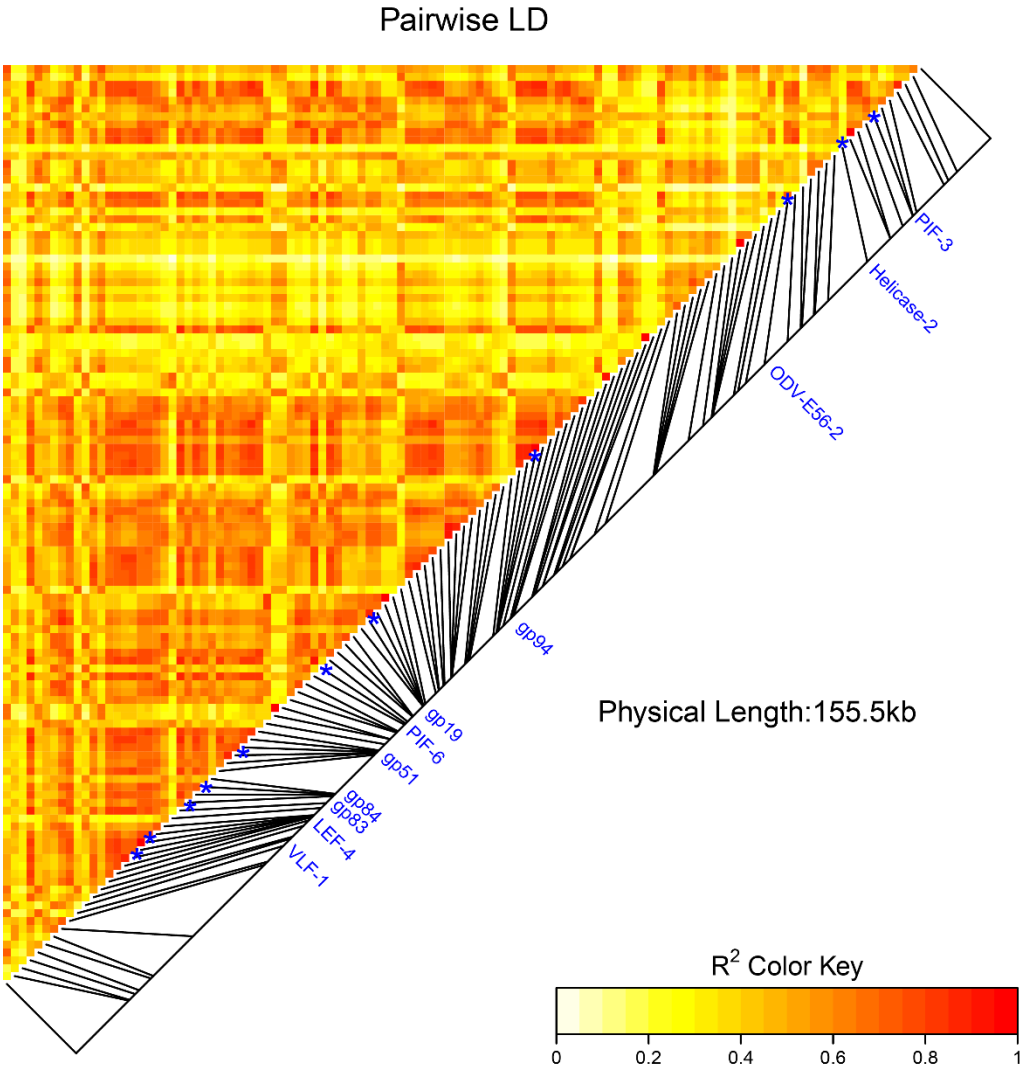
793

794 **Supplementary Figure 2:** Frequency of each significant SNP within each sample, ranked by the viral titer
795 in each sample (left = lowest, right = highest), to show the strong linkage of SNPs and little evidence of co-
796 infection, also highlights the association between SNP frequency and Helicase-2.

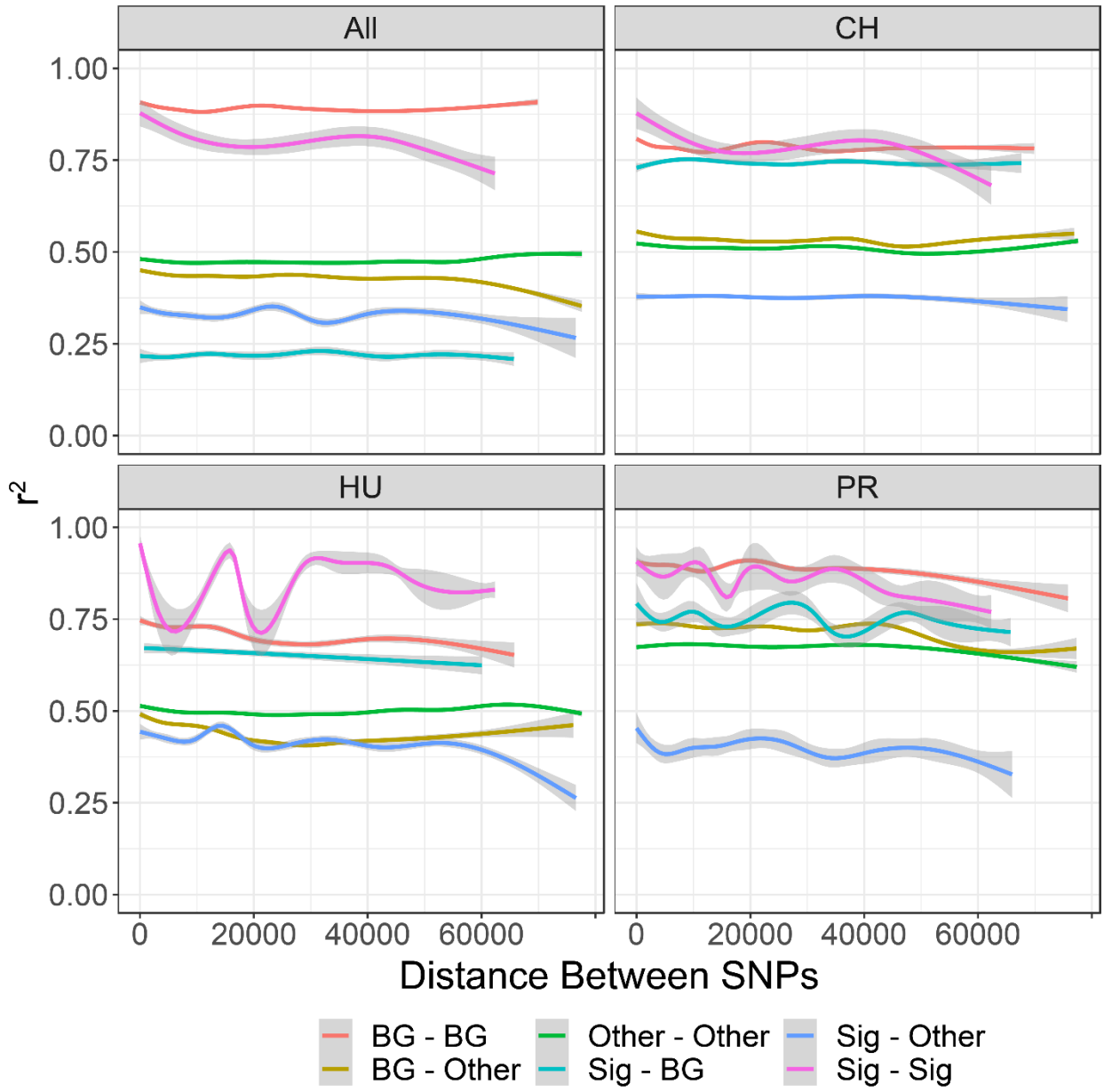


797

Supplementary Figure 3: Linkage disequilibrium between SNPs in DiNV. The labelled SNPs (significant SNPs found in the GWAS) are strongly linked. Points are colored by the estimated linkage between SNPs, from red (highly linked, $r^2 = 1$) to white (unlinked, $r^2 = 0$)

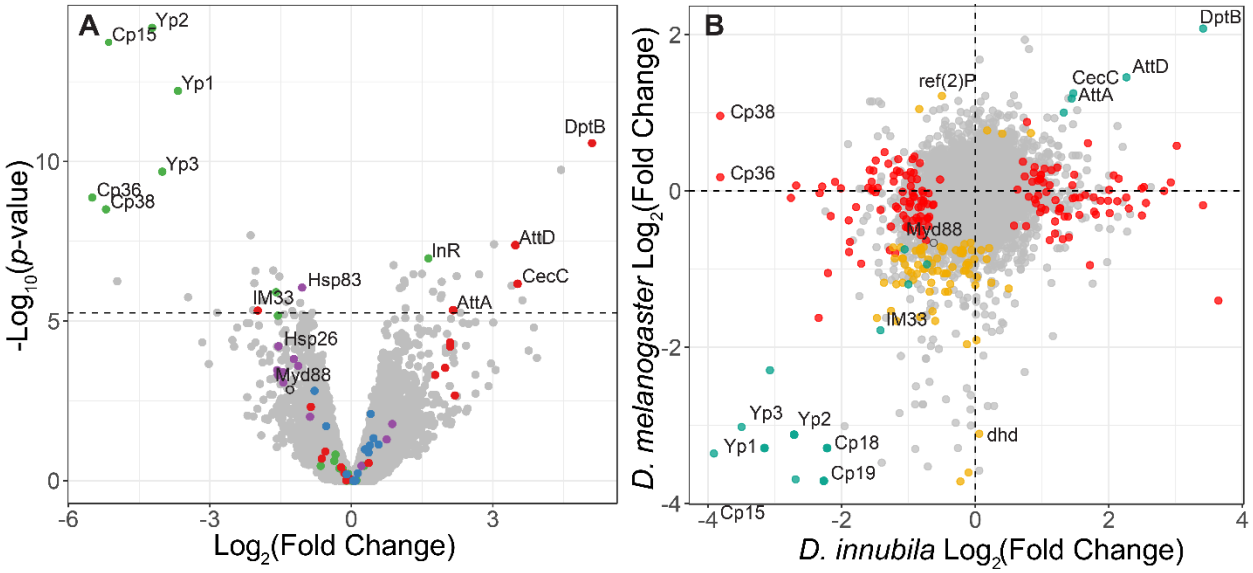


802 **Supplementary Figure 4:** Linkage (r^2) between different types of SNPs in each population of DiNV, and
803 across all samples. Other = SNPs which are not significantly associated with DiNV titer and do not form
804 the viral haplotype. Sig = SNPs which are significantly associated with DiNV titer and do not form the
805 viral haplotype. BG = SNPs which are in the background which the viral haplotype evolved on in each
806 population.



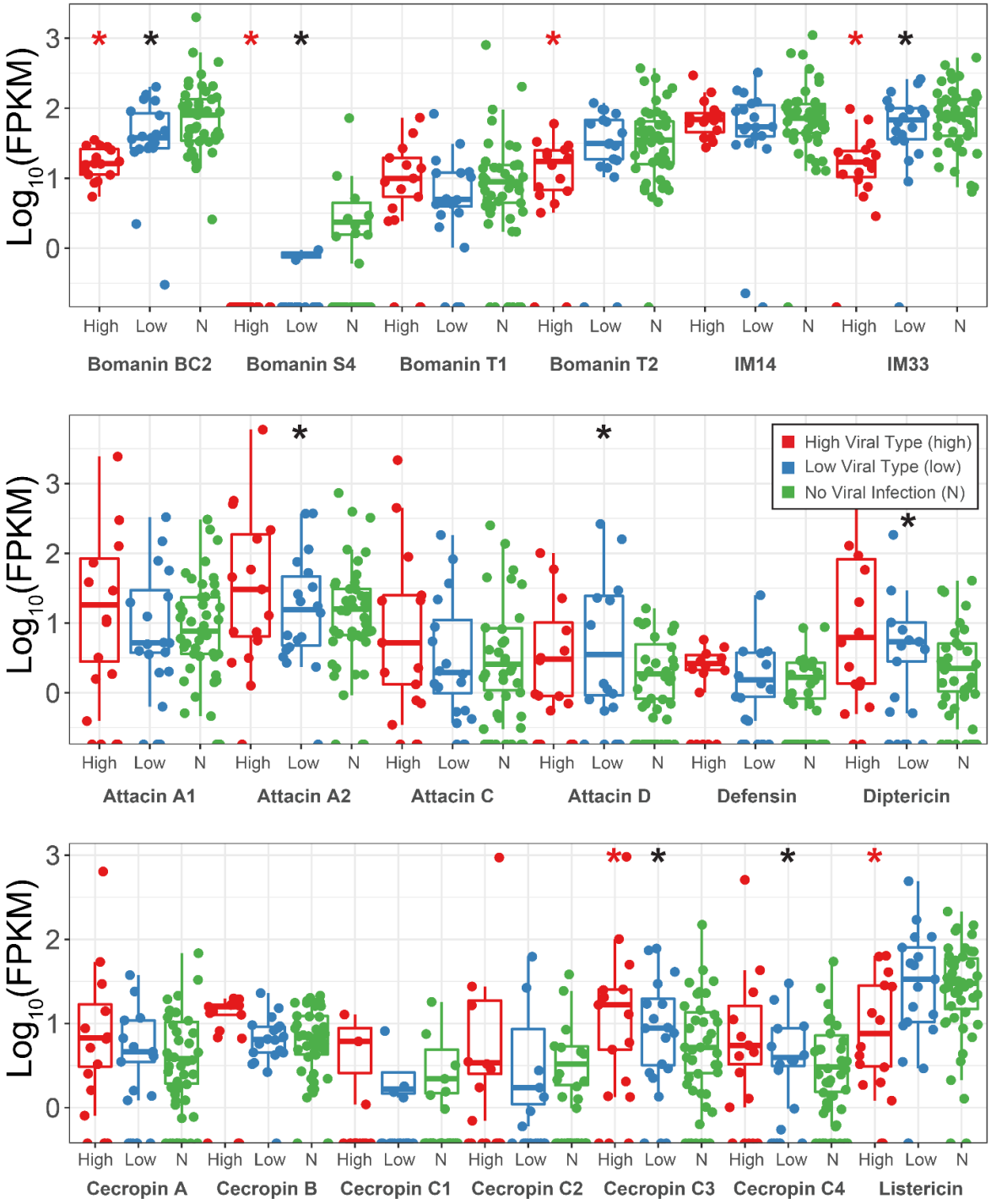
807

808 **Supplementary Figure 5:** Volcano plot of changes in gene expression between *D. innubila* infected with
809 DiNV and uninfected controls. Gene categories of interest, such as enriched categories, are highlighted in
810 color. The FDR-correct significance cut-off of 0.01 (10,320 tests) is shown as a dashed line. **B.** Comparison
811 of gene expression changes upon infection for *D. innubila* and *D. melanogaster*. Significantly differentially
812 expressed genes (p-value < 0.01, FDR-corrected) are colored, genes differentially expressed in both species
813 are colored blue, genes differentially expressed in just *D. melanogaster* are colored yellow and genes
814 differentially expressed in just *D. innubila* are colored red.

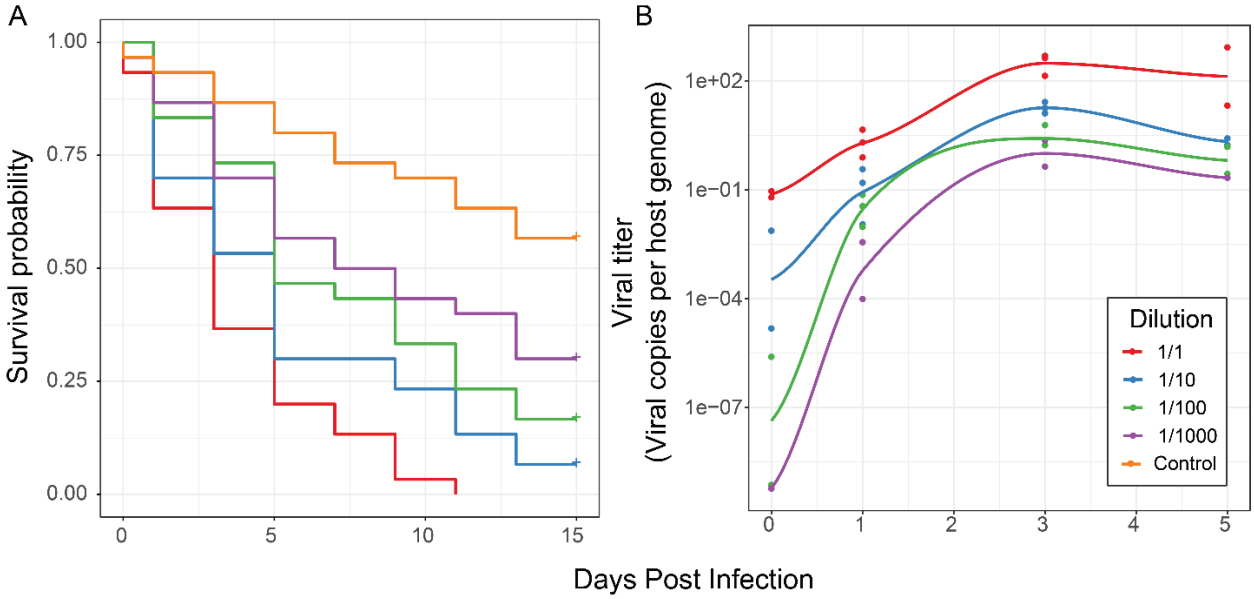


815

Supplementary Figure 6: Expression changes (shown as transcript fragments per 1 million reads per 1kbp of exon) of antimicrobial peptides between strains infected with high type DiNV, low type DiNV or not infected. Black stars above low samples show significant differential expression between DiNV infected strains and uninfected strains (multiple testing corrected *p*-value < 0.05). Red stars above high samples show significant differential expression between low type infected strains and high type infected strains (multiple testing corrected *p*-value < 0.05).

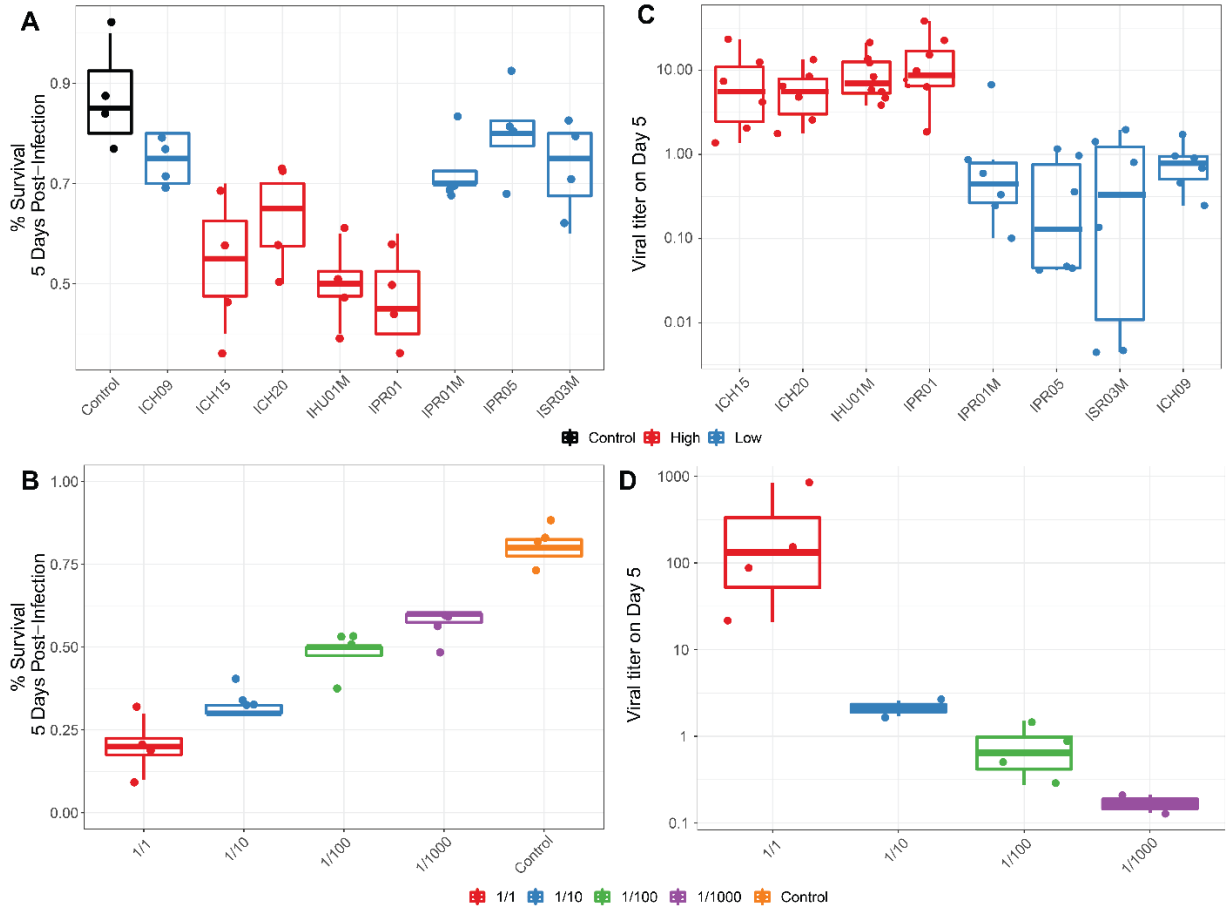


823 **Supplementary Figure 7:** Effect of differences in viral type and titer in experimental infections. **A.**
824 Survival curves of *D. innubila* infected with DiNV filtrate of different dilutions compared to control flies
825 pricked with sterile media, for 15 days post infection. **B.** qPCR copy number of viral *p47* relative to *tpi* in
826 samples of *D. innubila* infected with DiNV filtrate of different dilutions, between 1 and 1000 viral particles
827 per host genome copy.



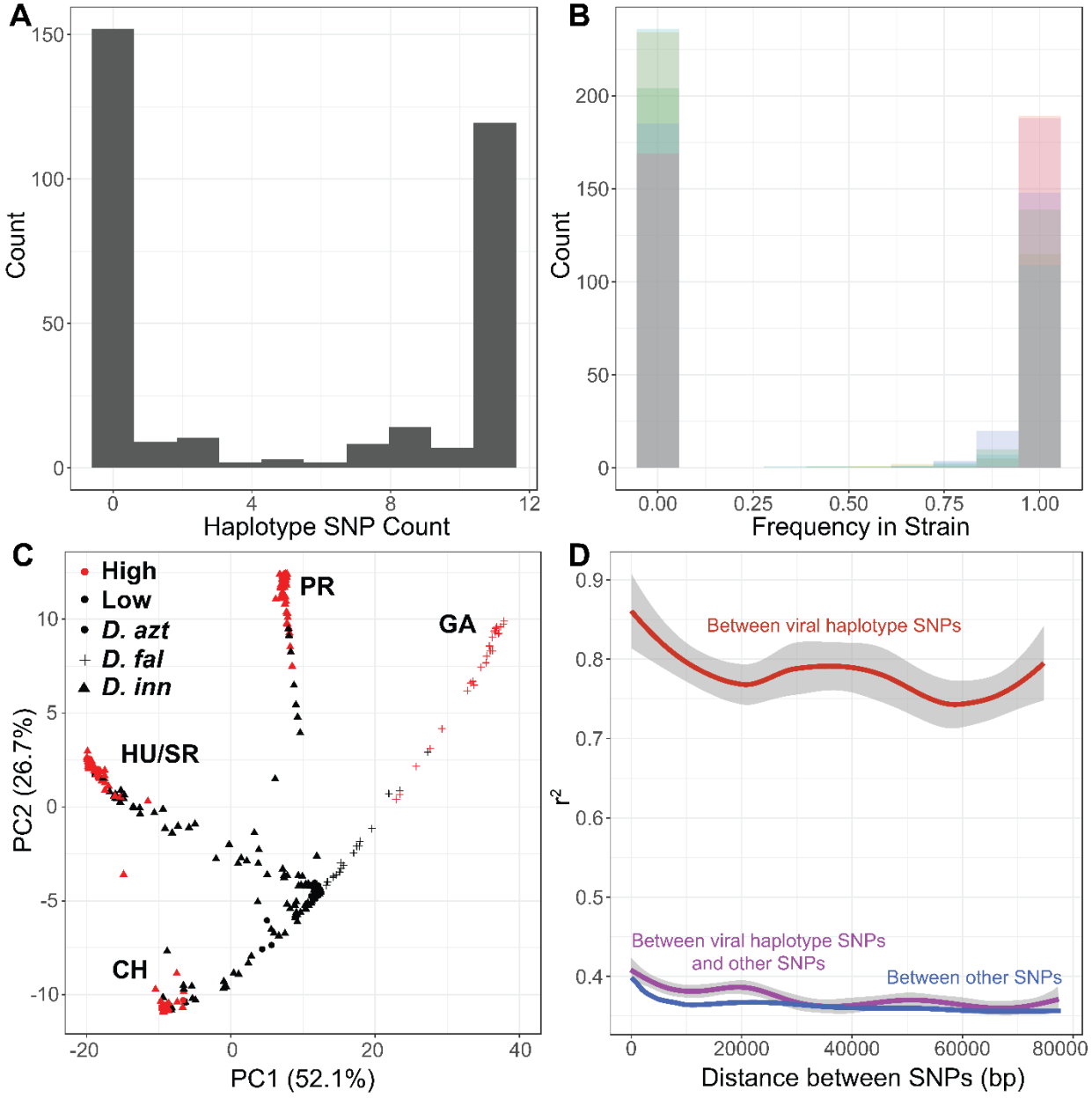
828

829 **Supplementary Figure 8 A:** Survival of *D. innubila* reference strain 5 days post infection, using filtrate
830 from different samples versus uninfected control, colored by high type virus or low type virus. **B.** Survival
831 of *D. innubila* reference strain 5 days post infection using serial dilutions of IPR01 filtrate versus control.
832 **C.** Viral titer estimated per viral genotype at 5 days post-infection, colored by high type virus or low type
833 virus. **D.** Viral titer of DiNV infecting *D. innubila* reference strain 5 days post infection using serial
834 dilutions of IPR01 filtrate versus control.



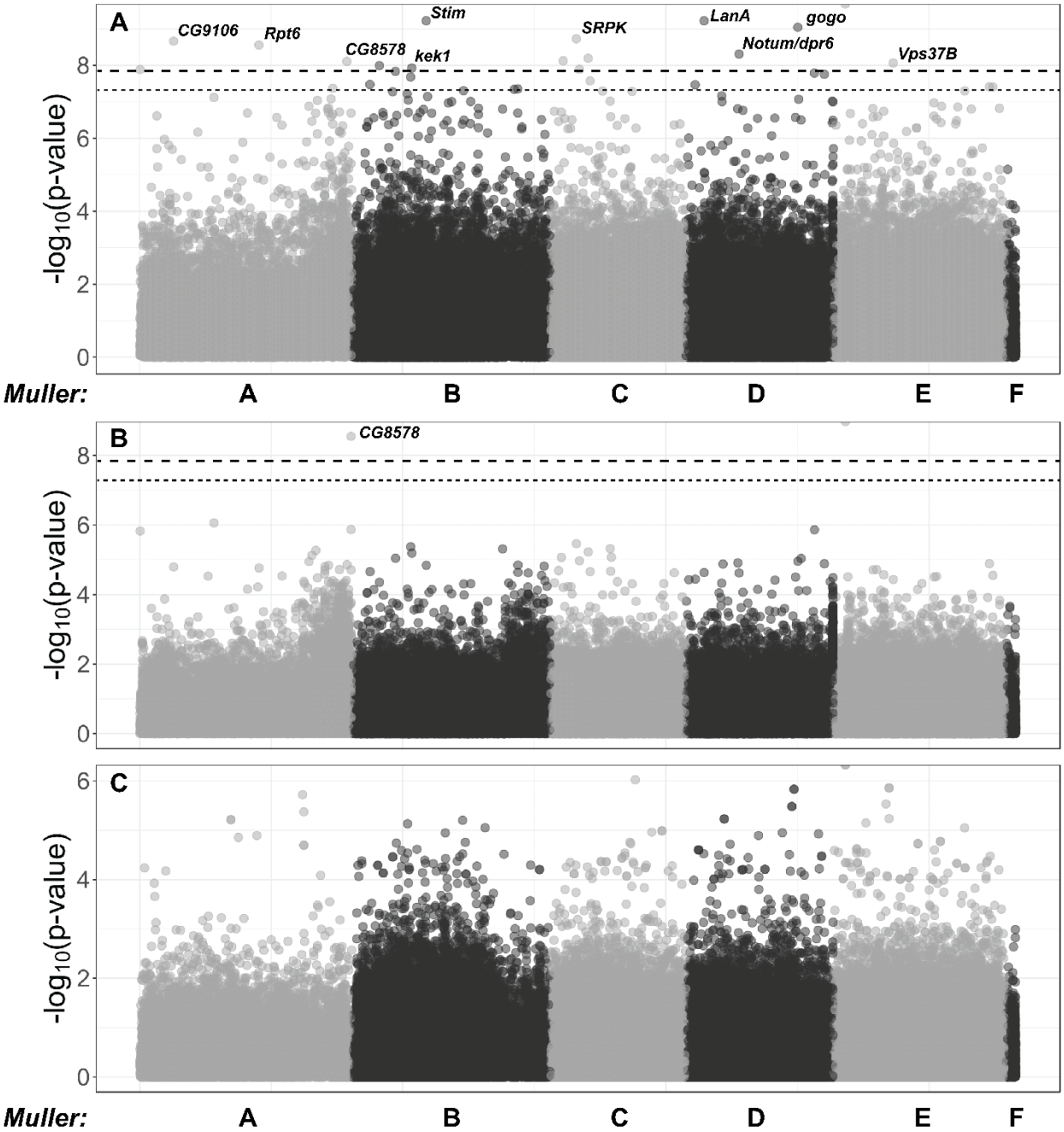
835

836 **Supplementary Figure 9. A**. Frequency of samples with different numbers of SNPs in the viral
837 haplotype, there ae very few intermediate types. **B.** Frequency of each SNP in samples infected with the
838 virus, showing there is little evidence of co-infections. **C.** Principle component analysis of DiNV strains
839 using variation of strains. Strains are colored by the viral type, showing its recurrent evolution. Point
840 shape denotes species in which DiNV was found (*D. azteca, D. falleni or D. innubila*). Strains cluster by
841 collection location. **D.** Linkage between SNPs in the viral haplotype (r^2) and other SNPs in the haplotype,
842 to other SNPs in the viral genome.



843

844 **Supplementary Figure 10. A:** Manhattan plot of significance of SNP on viral titer after factoring in
845 interaction with the viral haplotype. The significance cut offs are labelled (p-value < 0.05 after multiple
846 testing correction dotted, p-value < 0.05 after permutations dashed). **B**: Manhattan plot of SNP x viral
847 haplotype interaction for viral titer GWAS in *D. innubila*, calculated using *PLINK*. The significance cut
848 offs are labelled (p-value < 0.05 after multiple testing correction dotted, p-value < 0.05 after permutations
849 dashed). **C:** Manhattan plot of SNP x sex interaction for viral titer GWAS in *D. innubila*, calculated using
850 *PLINK*.



851

852 **Supplementary Figure 11:** Effective population size backwards for each population of DiNV going
853 backwards in time, estimated using StairwayPlot. Dotted lines indicate the error windows for N_e at a given
854 time point. Lines are colored by population.

855

856 **Supplementary Tables**

857 **Supplementary Table 1:** Summary of *Drosophila innubila* and *D. azteca* fly samples collected and
858 sequenced for this study, table includes summary of coverage for X chromosome, , Muller B, other
859 autosomes, virus and *Wolbachia*. Also contains SRA accessions for each strain.

860 **Supplementary Table 2:** Summary of *Drosophila innubila* fly RNA and DNA collected and sequenced
861 for this study, including if infected with DiNV.

862 **Supplementary Data**

863 **Supplementary Data 1:** FPKM of each gene in each sample across the whole *D. innubila* genome,
864 formatted for use in fitting a generalized linear model. Table include the gene name, gene flybase
865 annotation, *D. innubila* name, if the strains is infected with DiNV and the FPKM.

866 **Supplementary Data 2:** FPKM of each gene in each sample across the whole *D. innubila* genome,
867 formatted for differential gene expression analysis. Table include the gene name, gene flybase annotation,
868 *D. innubila* name, if the strains is infected with DiNV and the FPKM.

869 **Supplementary Data 3:** Differential gene expression analysis results between viral infected and
870 uninfected strains for both *D. innubila* and *D. melanogaster*. Genes are labelled as if differentially
871 expressed in one of the two species, or if differentially expressed in both.
872 **Supplementary Data 4:** VCF file for SNPs in DiNV, used in estimation of population genetic statistics
873 and in GWAS using PLINK.
874 **Supplementary Data 5:** VCF file for SNPs in *D. innubila*, used in estimation of population genetic
875 statistics and in GWAS.
876 **Supplementary Data 6:** Population genetic statistics calculated for each gene in *D. innubila* using
877 VCFtools for each population.
878 **Supplementary Data 7:** McDonald-Kreitman statistics calculated for each gene in *D. innubila* using
879 VCFtools for each population.
880 **Supplementary Data 8:** Population genetic statistics calculated for each gene in DiNV using VCFtools
881 for each population.
882 **Supplementary Data 9:** McDonald-Kreitman statistics calculated for each gene in DiNV using VCFtools
883 for each population.

884 **References**

885 Afonso, C. L., E. R. Tulman, Z. Lu, C. a. Balinsky, B. a. Moser *et al.*, 2001 Genome sequence of a
886 baculovirus pathogenic for *Culex nigripalpus.* Journal of virology 75**:** 11157-11165.
887 Alizon, S., and M. van Baalen, 2008 Transmission-virulence trade-offs in vector-borne diseases. Theor
888 Popul Biol 74**:** 6-15.
889 Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman, 1990 Basic local alignment search tool.
890 Journal of Molecular Biology 215**:** 403-410.
891 Anders, S., P. T. Pyl and W. Huber, 2015 HTSeq-A Python framework to work with high-throughput
892 sequencing data. Bioinformatics 31**:** 166-169.
893 Anderson, R. M., and R. M. May, 1982 Coevolution of hosts and parasites. Parasitology 85**:** 411-426.
894 Argaud, O., L. Croizier, M. López-Ferber and G. Croizier, 1998 Two key mutations in the host-range
895 specificity domain of the p143 gene of Autographa californica nucleopolyhedrovirus are required
896 to kill *Bombyx mori* larvae. Journal of General Virology 79**:** 931-935.
897 Blissard, G. W., and G. F. Rohrmann, 1990 Baculovirus diversity and molecular biology. Annual Review
898 of Entomology 35**:** 127-155.
899 Bouckaert, R., J. Heled, D. Kühnert, T. Vaughan, C. H. Wu *et al.*, 2014 BEAST 2: A Software Platform
900 for Bayesian Evolutionary Analysis. PLoS Computational Biology 10**:** 1-6.
901 Buffalo, V., 2018 *Scythe*.
902 Burgyan, J., and Z. Havelda, 2011 Viral suppressors of RNA silencing. Trends Plant Sci 16**:** 265-272.
903 Chateigner, A., A. Bezier, C. Labrousse, D. Jiolle, V. Barbe *et al.*, 2015 Ultra deep sequencing of a
904 baculovirus population reveals widespread genomic variations. Viruses 7**:** 3625-3646.
905 Cingolani, P., A. Platts, L. L. Wang, M. Coon, T. Nguyen *et al.*, 2012 A program for annotating and
906 predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of
907 *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly 6**:** 80-92.
908 Coccia, E. M., M. Severa, E. Giacomini, D. Monneron, M. E. Remoli *et al.*, 2004 Viral infection and Toll-
909 like receptor agonists induce a differential expression of type I and lambda interferons in human
910 plasmacytoid and monocyte-derived dendritic cells. European journal of immunology 34**:** 796-805.

Costa, A., E. Jan, P. Sarnow and D. Schneider, 2009 The Imd pathway is involved in antiviral immune responses in Drosophila. PLoS ONE 4.

Croizier, G., L. Croizier, O. Argaud and D. Poudevigne, 1994 Extension of Autographa californica nuclear polyhedrosis virus host range by interspecific replacement of a short DNA sequence in the p143 helicase gene. Proceedings of the National Academy of Sciences of the United States of America 91**:** 48-52.

Daugherty, M. D., and H. S. Malik, 2012 Rules of engagement: molecular insights from host-virus arms races. Annual review of genetics 46**:** 677-700.

Davey, N. E., G. Trave and T. J. Gibson, 2011 How viruses hijack cell regulation. Trends Biochem Sci 36**:** 159-169.

Dawkins, R., and J. R. Krebs, 1979 Arms Races between and within Species. Proceedings of the Royal Society of London B: Biological Sciences 205**:** 489-511.

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire *et al.*, 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics 43**:** 491-498.

Dobzhansky, T., 1937 Genetics and the origin of species. 364.

Dolan, P. T., Z. J. Whitfield and R. Andino, 2018 Mechanisms and Concepts in RNA Virus Population Dynamics and Evolution. Annual Review of Virology 5**:** 69-92.

Dyer, K. a., and J. Jaenike, 2005 Evolutionary dynamics of a spatially structured host-parasite association: *Drosophila innubila* and male-killing *Wolbachia*. Evolution; international journal of organic evolution 59**:** 1518-1528.

Eilertson, K. E., J. G. Booth and C. D. Bustamante, 2012 SnIPRE: Selection Inference Using a Poisson Random Effects Model. PLoS Computational Biology 8.

Enard, D., L. Cai, C. Gwennap and D. A. Petrov, 2016 Viruses are a dominant driver of protein adaptation in mammals. eLife 5**:** 1-25.

Feder, A. F., P. S. Pennings, J. Hermisson and D. A. Petrov, 2019 Evolutionary Dynamics in Structured Populations Under Strong Population Genetic Forces. G3 (Bethesda) 9**:** 3395-3407.

Ferreira, Á. G., H. Naylor, S. S. Esteves, I. S. Pais, N. E. Martins *et al.*, 2014a The Toll-Dorsal Pathway Is Required for Resistance to Viral Oral Infection in Drosophila. PLoS Pathogens 10.

Ferreira, Á. G., H. Naylor, S. S. Esteves, I. S. Pais, N. E. Martins *et al.*, 2014b The Toll-Dorsal Pathway Is Required for Resistance to Viral Oral Infection in *Drosophila.* PLoS Pathogens 10.

Gavrilets, S., 2003 Perspective: models of speciation: what have we learned in 40 years? Evolution; international journal of organic evolution 57**:** 2197-2215.

Gavrilets, S., 2004 Fitness Landscapes and the Origin of Species. 117-147.

Ghosh, S., and C. K. Chan, 2016 Analysis of RNA-Seq Data Using TopHat and Cufflinks. Methods Mol Biol 1374**:** 339-361.

Gifford, R. J., 2012 Viral evolution in deep time: lentiviruses and mammals. Trends Genet 28**:** 89-100.

Gillespie, J., 2004 Population Genetics: A Concise Guide. 232.

Haas, B. J., A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood *et al.*, 2013 De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature protocols 8**:** 1494-1512.

Hermisson, J., and P. S. Pennings, 2005 Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics 169**:** 2335-2352.

Hill, T., B. Koseva and R. L. Unckless, 2019 The genome of *Drosophila innubila* reveals lineage-specific patterns of selection in immune genes. Molecular Biology and Evolution**:** 1-36.

Hill, T., and R. L. Unckless, 2017 Baculovirus molecular evolution via gene turnover and recurrent positive selection of key genes. Journal of Virology 91**:** e01319-01317.

Hill, T., and R. L. Unckless, 2018 The dynamic evolution of Drosophila innubila Nudivirus. Infection, Genetics and Evolution**:** 1-25.

Hill, T., and R. L. Unckless, 2020 Selection and demography shape genomic variation in a 'Sky Island' species. Biorxiv**:** 1-35.

962 Hoffmann, J. A., 2003 The immune response of *Drosophila*. Nature 426**:** 33-38.
963 Holmes, E. C., 2007 Viral evolution in the genomic age. PLoS Biol 5**:** e278.
964 Http://broadinstitute.github.io/picard, *Picard*.
965 Hultmark, D., 2003 *Drosophila* immunity: Paths and patterns. Current Opinion in Immunology 15**:** 12-19.
966 Jackson, T., 2009 The Use of Oryctes Virus for Control of Rhinoceros Beetle in the Pacific Islands, pp.
967 133-140 in *Use of Microbes for Control and Eradication of Invasive Arthropods SE - 8*, edited by
968 A. Hajek, T. Glare and M. O'Callaghan. Springer Netherlands.
969 Jackson, T. A., A. M. Crawford and T. R. Glare, 2005 Oryctes virus - Time for a new look at a useful
970 biocontrol agent. Journal of Invertebrate Pathology 89**:** 91-94.
971 Joshi, N., and J. Fass, 2011 Sickle: A sliding window, adaptive, quality-based trimming tool for fastQ files.
972 1.33.
973 Kaltz, O., and J. A. Shykoff, 1998 Local adaptation in host–parasite systems. Heredity 81**:** 361-370.
974 Kassambara, A., M. Kosinski and P. Biecek, 2017 survminer: Drawing Survival Curves using'ggplot2'. R
975 package version 0.3 1.
976 Kelly, D. C., 1982 Baculovirus Replication. Journal of General Virology 63**:** 1-13.
977 Kondrashov, A. S., S. Sunyaev and F. A. Kondrashov, 2002 Dobzhansky–Muller incompatibilities in
978 protein evolution. PNAS 99**:** 14878-14883.
979 Lamiable, O., C. Kellenberger, C. Kemp, L. Troxler, N. Pelte *et al.*, 2016 Cytokine Diedel and a viral
980 homologue suppress the IMD pathway in *Drosophila*. Proceedings of the National Academy of
981 Sciences 113**:** 698-703.
982 Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows-Wheeler transform.
983 Bioinformatics (Oxford, England) 25**:** 1754-1760.
984 Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The sequence alignment/map format and
985 SAMtools. Bioinformatics (Oxford, England) 25**:** 2078-2079.
986 Lipsitch, M., S. Siller and M. Nowak, 1996 The evolution of virulence in pathogens with veritcal and
987 horizontal transmission. Evolution 50**:** 1729-1741.
988 Liu, X., and Y.-X. Fu, 2015 Exploring population size changes using SNP frequency spectra. Nature
989 genetics 47**:** 555-559.
990 Maeda, S., S. G. Kamita and A. Kondo, 1993 Host range expansion of Autographa californica nuclear
991 polyhedrosis virus (NPV) following recombination of a 0.6-kilobase-pair DNA fragment
992 originating from Bombyx mori NPV. Journal of virology 67**:** 6234-6238.
993 Martin, M., 2011 Cutadapt removes adapter sequences from high-throughput sequencing reads. Technical
994 Notes**:** 1-12.
995 May, R. M., and M. Nowak, 1995 Coinfection and the Evolution of Parasite Virulence. Proceedings:
996 Biological Sciences 261**:** 209-215.
997 McDonald, J. H., and M. Kreitman, 1991 Adaptive protein evolution at the Adh locus in *Drosophila*. Nature
998 351**:** 652-654.
999 McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.*, 2010 The Genome Analysis Toolkit:
1000 A MapReduce framework for analyzing next-generation DNA sequencing data. Proceedings of the
1001 International Conference on Intellectual Capital, Knowledge Management & Organizational
1002 Learning 20**:** 1297-1303.
1003 Merkling, S. H., and R. P. van Rij, 2013 Beyond RNAi: Antiviral defense strategies in Drosophila and
1004 mosquito. Journal of Insect Physiology 59**:** 159-170.
1005 Mukherjee, K., H. Campos and B. Kolaczkowski, 2013 Evolution of animal and plant dicers: early parallel
1006 duplications and recurrent adaptation of antiviral RNA binding in plants. Mol Biol Evol 30**:** 627-
1007 641.
1008 Nielsen, R., 2005 Molecular signatures of natural selection. Annu Rev Genet 39**:** 197-218.
1009 Nielsen, R., C. Bustamante, A. G. Clark, S. Glanowski, T. B. Sackton *et al.*, 2005 A scan for positively
1010 selected genes in the genomes of humans and chimpanzees. PLoS Biol 3**:** e170.

1011 Obbard, D. J., K. H. J. Gordon, A. H. Buck and F. M. Jiggins, 2009a The evolution of RNAi as a defence
1012 against viruses and transposable elements. Philosophical transactions of the Royal Society of
1013 London. Series B, Biological sciences 364: 99-115.
1014 Obbard, D. J., F. M. Jiggins, D. L. Halligan and T. J. Little, 2006 Natural Selection Drives Extremely Rapid
1015 Evolution in Antiviral RNAi Genes. Current biology 16: 580-585.
1016 Obbard, D. J., J. J. Welch, K. W. Kim and F. M. Jiggins, 2009b Quantifying adaptive evolution in the
1017 *Drosophila* immune system. PLoS Genetics 5: e1000698.
1018 Orr, H. A., 1995 Population genetics of speciation: the evolution of hybrid incompatibilities. Genetics 139:
1019 1805-1813.
1020 Orr, H. A., 2004 Dobzhansky, Bateson, and the genetics of speciation. Genetics 168: 1097-1104.
1021 Orr, H. A., and M. Turelli, 2001 The evolution of postzygotic isolation: accumulating Dobzhansky-Muller
1022 incompatibilities. Evolution; international journal of organic evolution 55: 1085-1094.
1023 Palmer, W. H., J. D. Hadfield and D. J. Obbard, 2018a RNA Interference Pathways Display High Rates of
1024 Adaptive Protein Evolution in Multiple Invertebrates. Genetics: genetics.300567.302017.
1025 Palmer, W. H., J. Joosten, G. J. Overheul, P. W. Jansen, M. Vermeulen *et al.*, 2019 Induction and
1026 suppression of NF- κB signalling by a DNA virus of *Drosophila*. Journal of Virology 93: e01443-
1027 01418.
1028 Palmer, W. H., N. Medd, P. M. Beard and D. J. Obbard, 2018b Isolation of a natural DNA virus of
1029 Drosophila melanogaster, and characterisation of host resistance and immune responses. PLoS
1030 Pathogens 4: e1007050.
1031 Palmer, W. H., N. Medd, P. M. Beard and D. J. Obbard, 2018c Isolation of a natural DNA virus of
1032 *Drosophila melanogaster*, and characterisation of host resistance and immune responses. PLoS
1033 Pathogens 14: 1-26.
1034 Paradis, E., J. Claude and K. Strimmer, 2004 APE: analyses of phylogenetics and evolution in R language.
1035 Bioinformatics 20: 289-290.
1036 Pennings, P. S., 2012 Standing genetic variation and the evolution of drug resistance in HIV. PLoS
1037 Computational Biology 8.
1038 Pennings, P. S., S. Kryazhimskiy and J. Wakeley, 2014 Loss and Recovery of Genetic Diversity in Adapting
1039 Populations of HIV. PLoS Genetics 10.
1040 Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira *et al.*, 2007 PLINK: A Tool Set for
1041 Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of
1042 Human Genetics 81: 559-575.
1043 Quast, C., E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer *et al.*, 2013 The SILVA ribosomal RNA gene
1044 database project: improved data processing and web-based tools. Nucleic Acids Res 41: D590-596.
1045 Robinson, M. D., D. J. McCarthy and G. K. Smyth, 2009 edgeR: A Bioconductor package for differential
1046 expression analysis of digital gene expression data. Bioinformatics 26: 139-140.
1047 Rohrmann, G. F., 2013 Baculovirus Molecular Biology. 211.
1048 Sackton, T. B., B. P. Lazzaro, T. A. Schlenke, J. D. Evans, D. Hultmark *et al.*, 2007 Dynamic evolution of
1049 the innate immune system in *Drosophila*. Nature Genetics 39: 1461-1468.
1050 Sagulenko, P., V. Puller and R. A. Neher, 2018 TreeTime: Maximum-likelihood phylodynamic analysis.
1051 Virus Evol 4: vex042.
1052 Schulz, M. H., D. R. Zerbino, M. Vingron and E. Birney, 2012 Oases: Robust de novo RNA-seq assembly
1053 across the dynamic range of expression levels. Bioinformatics 28: 1086-1092.
1054 Shultz, A., and T. B. Sackton, 2019 Immune genes are hotspots of shared positive selection across birds
1055 and mammals. eLife 8: e41815.
1056 Smit, A. F. A., and R. Hubley, 2008 RepeatModeler Open-1.0.
1057 Smit, A. F. A., and R. Hubley, 2013-2015 RepeatMasker Open-4.0, pp. RepeatMasker.
1058 Soetaert, K., T. Petzoldt and R. W. Setzer, 2010 Solving Differential Equations in R: Package deSolve.
1059 Journal of Statistical Software 33: 1-24.
1060 Stoletzki, N., and A. Eyre-Walker, 2011 Estimation of the neutrality index. Molecular Biology and
1061 Evolution 28: 63-70.

1062 Team, R. C., 2013 R: A Language and Environment for Statistical Computing, pp. R Foundation for
1063 Statistical Computing, Vienna, Austria.
1064 Unckless, R. L., 2011 A DNA Virus of *Drosophila*. PLoS ONE 6**:** e26564.
1065 Wang, X.-H. X.-H., R. Aliyari, W.-X. Li, H.-W. Li, K. Kim *et al.*, 2006 RNA interference directs innate
1066 immunity against viruses in adult *Drosophila*. Science 312**:** 452-454.
1067 Wang, Y., and J. a. Jehle, 2009 Nudiviruses and other large, double-stranded circular DNA viruses of
1068 invertebrates: new insights on an old topic. Journal of invertebrate pathology 101**:** 187-193.
1069 Wang, Y., R. G. Kleespies, M. B. Ramle and J. A. Jehle, 2008 Sequencing of the large dsDNA genome of
1070 Oryctes rhinoceros nudivirus using multiple displacement amplification of nanogram amounts of
1071 virus DNA. Journal of Virological Methods 152**:** 106-108.
1072 Webster, C. L., F. M. Waldron, S. Robertson, D. Crowson, G. Ferrari *et al.*, 2015 The discovery,
1073 distribution, and evolution of viruses associated with *Drosophila melanogaster.* PLoS biology 13**:**
1074 e1002210.
1075 West, C., and N. Silverman, 2018 p38b and JAK-STAT signaling protect against invertebrate iridescent
1076 virus 6 infection in *Drosophila*. Plant Pathology 14**:** e1007020.
1077 Williams, G. C., and R. M. Nesse, 1991 The Dawn of Darwinian Medicine. The Quarterly Review of
1078 Biology 66**:** 1-22.
1079 Wilm, A., P. Poh, K. Aw, D. Bertrand, G. Hui *et al.*, 2012 LoFreq: a sequence-quality aware, ultra-sensitive
1080 variant caller for uncovering cell-population heterogeneity from high-throughput sequencing
1081 datasets. Nucleic Acids Research 40**:** 11189-11201.
1082 Wilson, B. A., N. R. Garud, A. F. Feder, Z. J. Assaf and P. S. Pennings, 2016 The population genetics of
1083 drug resistance evolution in natural populations of viral, bacterial and eukaryotic pathogens. Mol
1084 Ecol 25**:** 42-66.
1085 Wu, T. D., and S. Nacu, 2010 Fast and SNP-tolerant detection of complex variants and splicing in short
1086 reads. Bioinformatics (Oxford, England) 26**:** 873-881.
1087 Zambon, R. A., M. Nandakumar, V. N. Vakharia and L. P. Wu, 2005a The Toll pathway is important for
1088 an antiviral response in *Drosophila*. Proceedings of the National Academy of Sciences 102**:** 7257-
1089 7262.
1090 Zambon, R. A., M. Nandakumar, V. N. Vakharia and L. P. Wu, 2005b The Toll pathway is important for
1091 an antiviral response in *Drosophila*. Proceedings of the National Academy of Sciences 102**:** 7257-
1092 7262.
1093