Long title: Host evolutionary history predicts virus prevalence across bumblebee species

Short title: Evolutionary signals in bumblebee-virus interaction networks

David J. Pascall[1,6], Matthew C. Tinsley[2], Darren J. Obbard[3,4] and Lena Wilfert[1,5]

[1] College of Life and Environmental Science, University of Exeter Cornwall Campus, Treliever Road, Penryn, UK

[2]

[3] Institute of Evolutionary Biology, University of Edinburgh, Charlotte Auerbach Road, Edinburgh, UK

[4] Centre for Infection, Evolution and Immunity, University of Edinburgh, Charlotte Auerbach Road, Edinburgh, UK

[5] Institute of Evolutionary Ecology and Conservation Genomics, University of Ulm, Albert-Einstein-Allee 11, 89069 Ulm, Germany

[6] Institute of Biodiversity, Animal Health and Comparative Medicine, Graham Kerr Building, University of Glasgow, Glasgow, UK

Author Contributions

DJP – Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

MCT – Investigation, Resources, Writing – Review & Editing

DJO – Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Methodology, Project Administration, Software, Visualisation, Writing – Review & Editing

LW – Conceptualization, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Supervision, Writing – Review & Editing

Author's Summary

Despite the importance of disease in the regulation of animal populations, our understanding of the distribution of pathogen burden across wild communities remains in its infancy. In this study, we investigated the distribution of viruses across natural populations of 13 different bumblebee species in Scotland. In order to accurately assess this distribution, we first searched for viruses using a transcriptomic approach, finding at least 30 new viruses of bumblebees, and assayed a subset of them for their presence and absence in different species. Then, in the first application of these methods to an animal-virus system, we used co-phylogenetic mixed models to investigate the factors that lead to species being infected to different degrees by a subset of these viruses. We found that, while much of the variation in the prevalence of the viruses can be explained by the specifics of individual bumblebee-virus pairings, related bumblebee species being infected to similar degrees with the same sets of viruses has an important contribution to the distribution of viruses across hosts. Consistent with previous work, our study indicates that, while in general the interaction between a host and a virus may be unpredictable, closely related species are more likely to exhibit similar patterns.

Abstract

Why a pathogen associates with one host but not another is one of the most important questions in disease ecology. Here we use transcriptome sequencing of wild-caught bumblebees from 13 species to describe their natural viruses, and to quantify the impact of evolutionary history on the realised associations between viruses and their pollinator hosts. We present 37 novel virus sequences representing at least 30 different viruses associated with bumblebees. We verified 17 of them by PCR and estimate their prevalence across species in the wild. Through small RNA sequencing, we demonstrate that at least 10 of these viruses form active infections in wild individuals. Using a phylogenetic mixed model approach, we show that the evolutionary history of the host shapes the current distribution of virus/bumblebee associations. Specifically, we find that related hosts share viral assemblages, viruses differ in their prevalence averaged across hosts and the prevalence of infection in individual virus-host pairings depends on precise characteristics of that pairing.

Introduction

Pathogens that naturally infect more than one host species have a particularly high risk of disease emergence (Woolhouse & Gowtage-Sequeria 2005). One especially important group of pathogens are the viruses, whose ubiquity leads them to have a disproportionate role in the regulation of natural populations (Suttle 2007). Viruses are relevant in populations that humans manage for economic and conservation reasons, such as bumblebees, which are both in decline (Williams & Osborne 2009) and important providers of ecosystem services (Garibaldi et al. 2013).

75 Bumblebees, genus *Bombus*, are a primitively eusocial group of important wild
76 pollinators; many bumblebee species have experienced population declines, linked to
77 biotic and abiotic stressors such as habitat degradation, pesticide use and shared
78 infectious diseases for example caused by viral pathogens (Vanbergen & the Insect
79 Pollinators Initiative 2013). While honeybee viruses have been intensively studied,
80 and have in many cases been found to represent multihost pathogens (see Manley et
81 al. (2015) and the references within), bumblebee-specific viruses are comparatively
82 poorly studied, and how widely they are shared between species is unknown.
83
84 In order for a species to be a multihost pathogen, some degree of opportunity for
85 cross-species transmission must exist. Our definition of multihost pathogens follows
86 that of Fenton et al. (2015). As such, multihost pathogens are defined to include two
87 conceptually distinct groups: 'facultative multihost pathogens' that are able to
88 maintain transmission chains in multiple host species (i.e. R_0>=1 in multiple host
89 species) and 'obligate multihost pathogens', which rely on sufficiently high rates of
90 cross-species transmission to offset unsustainable transmission within individual host
91 species (i.e. 0<R_0<1 within host species, R_0>=1 overall). In addition, pathogens that
92 maintain transmission in a single host (R_0>=1) but experience regular spillover (with
93 or without the expectation of onward transmission: 0<= R_0<1) are included as being
94 effectively multihost pathogens within our definition. R_0 is defined as the expected
95 number of secondary infections caused by a single typical infected individual in an
96 entirely naïve host population (Heesterbeek 2002). We define cross-species
97 transmission as the movement of a multihost pathogen between host species within its
98 host range. This contrasts with host shifting, which we define as a transmission event
99 to a new host species, leading to a change in host range; however there is necessarily

100 some unavoidable ambiguity between cross-species transmission and host shifting in
101 the case of pathogens that exhibit rare spillover events.
102
103 The opportunity for cross-species transmission, which explains the large number of
104 viruses originally detected in honeybees present in bumblebees, may be created by
105 niche overlap in foraging (Salathé & Schmid-Hempel 2011). Bumblebee nests are
106 provisioned by foraging workers who gather pollen and nectar from flowers in the
107 surrounding area. Considerable interspecific differences in plant species utilization by
108 foragers of different species are commonly observed (e.g. Arbulo, Santos, Salvarrey
109 & Invernizzi 2011; Goulson & Darvill 2004; Goulson et al. 2008; Harder 1985), but
110 this is not a universal phenomenon (Lye et al. 2010), and the degree of overlap may
111 depend on the diversity of flowers currently in bloom. Flower choice of foragers is
112 correlated with species tongue length (Goulson et al. 2008; Harder 1985), which
113 implicitly incorporates shared behavioural characteristics between closely related
114 bumblebee species as there is phylogenetic correlation between tongue length and
115 relatedness (Harmon-Threatt & Ackerly 2013). Different species of bumblebee also
116 exhibit incomplete temporal separation throughout the year, causing some degree of
117 partitioning in niche space even when they are spatially sympatric (Goodwin 1995).
118 This ecology leads to a complex interaction network between bumblebee species as
119 well as sympatric honeybees, which may structure cross-species transmission.
120
121 The prevalence of pathogens, including viruses, across host species, such as
122 bumblebees, is structured on two levels. First, a virus may be present or entirely
123 absent in a potential host species. Second, other factors may then influence how
124 prevalent a pathogen is within that species. At the presence/absence level, a complete

125 lack of infection in nature can occur in three ways: 1) a host and virus may exist in
126 allopatry or in completely non-interacting ecological niches, preventing transmission
127 irrespective of the host's susceptibility; 2) a physiological or molecular mismatch
128 (including immunity) between a host and virus can prevent infection; and 3)
129 environmental conditions may be such that transmission cannot occur between two
130 sympatric species. None of these mechanisms represent an immutable barrier, and all
131 represent ends of a continuum, where lesser forms simply reduce transmission.
132 Spatially or ecologically separated hosts and parasites may come into contact through
133 migrations or human facilitated invasions, allowing new associations to emerge. For
134 example, the arrival of *Plasmodium relictum* to the Hawaiian islands led to avian
135 population declines and contributed to extinctions in the naturally susceptible but
136 naïve populations (van Riper et al. 1986). Incompatibility can break down if evolution
137 in the pathogen or host removes the physiological or molecular barriers to infection,
138 as shown when Canine parvovirus type 2 emerged from Feline panleukopenia virus
139 after gaining the ability to bind to canine transferrin receptors (Hueffer et al. 2003).
140
141 For virus-host associations where infection can and does occur, quantitative
142 differences in infection risk between species can be driven by ecological variation in
143 transmission rates. These differences can be driven by, for example, the propensity for
144 group living (Johnson et al. 2011), population densities (Arneberg et al. 1998), the
145 biodiversity of the community (Civitello et al. 2015) and host avoidance behaviours
146 (Curtis 2014). Variation in infection risk among host species can also be driven by
147 physiological and molecular factors, with hosts having varying suitability for the
148 replication of a given parasite. In the extreme case, a host species may exhibit
149 condition-dependent susceptibility; where infection can only occur when the immune

150 system is suppressed, either directly, through an immunosuppressant disease or
151 chemical agent, or indirectly, through trade-offs in resource allocation brought about
152 by malnutrition (Chandra 1983). Both behavioral and ecological factors, leading to
153 differences in contact rate, and physiological factors, leading to differences in
154 infection probability on contact, may be phylogenetically correlated (Harmon-Threatt
155 & Ackerly 2013; Longdon et al. 2011).
156
157 **-------**
158 **Box 1 – Definition of Terms**
159 Co-phylogenetic generalized linear mixed models that incorporate phylogenetic
160 variance from multiple clades (Hadfield et al. 2014; Rafferty & Ives 2013) have been
161 used relatively rarely, and a biological interpretation of the model terms may not be
162 immediately familiar. In the host-parasite context, this approach can be used to
163 model how the probability of infection is predicted by both host and parasite species,
164 allowing for covariance induced by the relationships within each group, and the
165 interactions between these model terms. This can be considered either at the species-
166 wide level (i.e. the probability that infection will occur at all in a given host/parasite
167 pairing), or at the level of individuals within species (i.e. infection prevalence). Here
168 we provide verbal descriptions of how the terms can be interpreted, as well as
169 references to a figure in Hadfield et al. (2014) where each of these effects is
170 illustrated graphically:
171
172 *Phylogenetic Effect*: Variation in the mean value of a trait among species that is
173 explained by phylogenetic divergence. For example, more closely-related hosts might
174 be more similar in susceptibility to viral infection (display higher viral prevalence),

175 irrespective of virus species. Equivalently, more closely-related viruses might be
176 more similar in infectiousness, irrespective of host species (Figures 1a and b in
177 Hadfield et al. (2014)).
178
179 *Species Effect*: Variation in the mean value of a trait among species that is
180 not explained by a *Phylogenetic Effect*. For example, much of the variation in
181 prevalence among viral species (irrespective of host), may not be explained by the
182 virus phylogeny but instead depend on lineage-specific viral traits (Figures 1g and f in
183 Hadfield et al. (2014)).
184
185 *Non-phylogenetic Interaction*: The interaction term between host and parasite *Species*
186 *Effects*, such that variation in the mean value of a trait within particular host/parasite
187 pairings depends the specifics of the host and parasite involved in a way not affected
188 by their evolutionary divergence. For example, variation in prevalence between
189 particular pairings that is caused by the interaction between lineage-specific host and
190 viral traits (Figure 1h in Hadfield et al. (2014)).
191
192 *Coevolutionary Interaction*: The interaction term between host and parasite
193 *Phylogenetic Effects*, such that variation in the mean value of a trait within particular
194 host/parasite pairings depends on the evolutionary divergence among species in both
195 host and parasite clades. For example, if the prevalence of infection is more similar
196 among pairings of closely-related hosts and closely-related parasites than would be
197 expected from the host and parasite phylogenies and species-means alone. (Figure 1e
198 in Hadfield et al. (2014)).
199

Evolutionary Interaction: The interaction term between the host (or parasite) *Phylogenetic Effect* and the partners' *Species Effect*, such that the variation in the mean value of a trait within particular host/parasite pairings depends on the evolutionary divergence among hosts (or parasites) and on the identity of particular partner species, but is not predicted by the evolutionary divergence between partners species. For example, if the similarity in viral prevalence for one virus species is strongly predicted by the evolutionary divergence among hosts, but a completely different relationship (unrelated to the evolutionary divergence among viruses) is seen for other virus species. (Figure 1c and d in Hadfield et al. (2014)).

A new multihost virus can arise in two ways, either through a virus species gaining a new host (Longdon et al. 2014) (Figure 1a), or through a speciation event in a multihost virus (Figure 1b). When novel multihost viruses are generated through host shifting (Figure 1a), a 'host evolutionary interaction' effect (see Box1) can result, as the consistent switching of viruses (V) to hosts (H) closely related to their ancestral host will lead to related hosts having correlated viral assemblages. When novel multihost viruses arise through speciation, i.e. if the ability to infect multiple hosts is an ancestral trait (Figure 1b), a 'virus evolutionary interaction' effect can result (see Box 1) through the inheritance of the ancestral host range, leading to the daughter virus species having correlated host assemblages. These effects can also be generated in ecological time through mechanisms that lead to biased cross-species transmission.

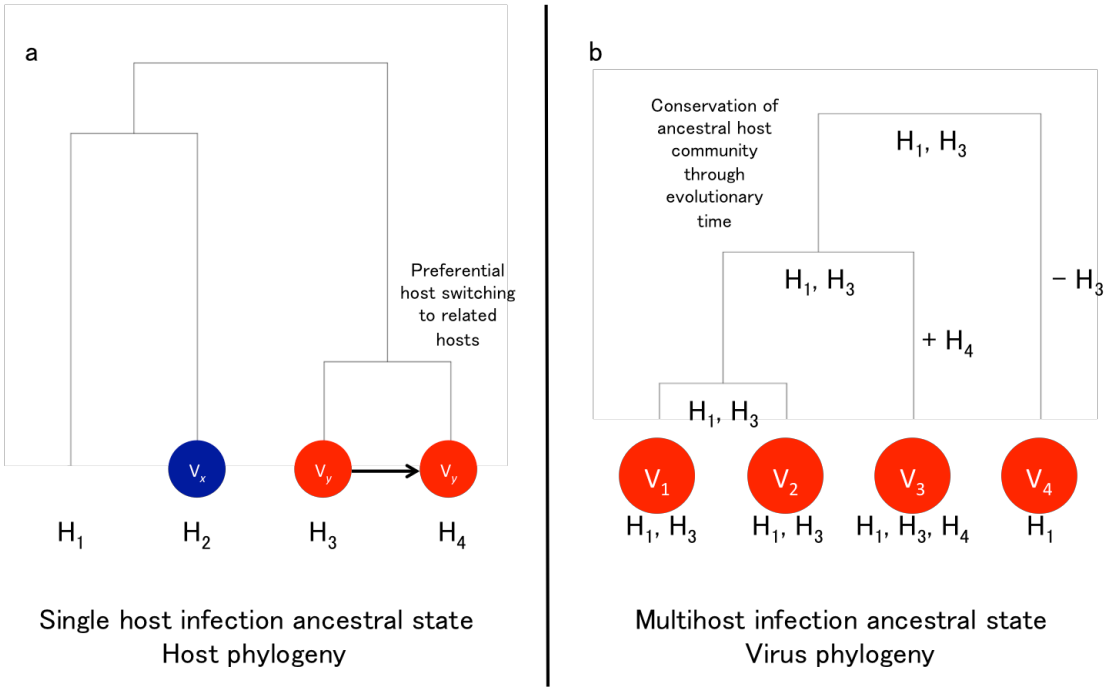


Figure 1 Mechanisms for the generation of novel multihost viruses. The generation of novel multihost viruses through host shifting (1a) leads to a 'host evolutionary interaction' effect (Box1), as the consistent switching of viruses (V) to hosts (H) closely related to their ancestral host will lead to related hosts having correlated viral assemblages. The generation of novel multihost viruses through speciation (1b) can lead to a 'virus evolutionary interaction' effect (Box 1) through the inheritance of the ancestral host range, leading to the daughter virus species having correlated host assemblages.

We tested for a role of evolutionary history in shaping the current host/virus assemblage using species from an ecologically and economically important group, the bumblebees. We cataloged the virome of wild-caught bumblebees from across Scotland by RNAseq, finding at least 30 new viruses. We then tested multiple bumblebee species for a subset of these novel viruses and three previously reported honeybee viruses: Slow bee paralysis virus (Bailey & Woods 1974), Acute bee paralysis virus (Bailey et al. 1963) and Hubei partiti-like virus 34 (Cornman et al. 2012; Shi et al. 2016). We analysed virus prevalence using co-phylogenetic models to determine the presence or absence and relative strengths of the evolutionary signals

241 that are expected to shape the host/virus assemblage in this system, and performed
242 tests to attempt to determine the mechanisms driving this.
243
244 **Methods**
245 **Sampling strategy**
246 A total of 926 individual bumblebees of 13 species were collected on the wing from
247 nine sites across Scotland in July and August of 2009 and 2011, and frozen in liquid
248 nitrogen or at -80°C. In 2009, we sampled the Ochil Hills, Glenmore, Dalwhinnie,
249 Stirling, Iona, Staffa, and the Pentlands, and in 2011 we sampled Edinburgh and
250 Gorebridge (Supplementary Table 1). The cryptic species complex of *Bombus*
251 *terrestris*, *Bombus lucorum, Bombus cryptarum* and *Bombus magnus* was resolved
252 using RFLP analysis following Murray et al. (2008). All individuals were bisected
253 longitudinally prior to RNA extraction. One half of each bumblebee was used in
254 pooled RNA extractions of 2-11 individuals per species (median 10; Supplementary
255 Table 2). Two of these pools ('DIV' and 'P11') were included in the RNAseq, but
256 excluded from prevalence testing. The groups of bumblebees were ground in liquid
257 nitrogen and added to TRIzol reagent (Life Technologies) for RNA extractions,
258 following the manufacturer's standard protocol. The RNA concentrations in the
259 pooled samples were equalized to approximately 200 ng/ul/individual based on
260 Nanodrop measurements.
261
262 **RNA Sequencing and Bioinformatics**
263 The RNA was combined by species for *B. terrestris* (239 individuals), *Bombus*
264 *pascuorum* (212 individuals), *B. lucorum* (182 individuals) and other *Bombus* (293
265 individuals) into four large RNA pools. These large pools were sequenced using the

266 Illumina HiSeq platform with 100bp paired end reads (Beijing Genomics Institute)
267 after poly-A selection. This excludes ribosomal and bacterial RNA, and will enrich
268 for mRNAs and those RNA viruses that have polyadenylated genomes or products.
269 The single-species bumblebee pools were subsequently re-sequenced following
270 duplex specific nuclease normalization, to reduce rRNA representation, and enrich for
271 rare transcripts while retaining non-polyadenylated viruses and products. The small
272 RNAs of the same RNA pools of *B. terrestris*, *B. lucorum* and *B. pascuorum* were
273 also sequenced to test for the replication of viruses identified via the transcriptome
274 sequencing.
275
276 For each pool, paired end RNAseq data were initially mapped to the published
277 *Bombus terrestris* and *B. impatiens* genomes using bowtie2 (Langmead & Salzberg
278 2012) to reduce the representation of conserved bumblebee sequences. Read pairs that
279 did not map concordantly, including divergent bumblebee sequences and other
280 associated microbiota, were assembled *de novo* using Trinity 2.2.0 (Grabherr et al.
281 2011) as paired end libraries, following automated trimming ('--trimmomatic') and
282 digital read normalisation ('--normalize_reads'). Where two RNAseq libraries (Poly-
283 A and DSN) had been sequenced, these were combined for assembly.
284
285 To identify putative viruses, all long open reading frames from each contig were
286 identified and concatenated to provide a 'bait' sequence for similarity searches using
287 Diamond (Buchfink et al. 2015) and BLASTp (Altschul et al. 1990). Contigs shorter
288 than 500 base pairs were discarded. These contig translations were used to search
289 against a Diamond database comprising all of the virus protein sequences available in
290 NCBI database 'nr', and all of the Dipteran, Hymenopteran, Nematode, Fungal,

291 Protist, and prokaryotic proteins available in NCBI database 'refseq_protein' (mode
292 'blastp'; e-value 0.001; maximum of one match). Matches to phage and short matches
293 to large DNA viruses were excluded. Remaining contigs were manually curated to
294 identify and annotate high-confidence virus-like sequences. To quantify approximate
295 fold-coverage, and to assess viRNA properties, the raw RNAseq and trimmed small
296 RNA reads were mapped against the putative viral contigs using bowtie2's '--very-
297 sensitive' setting and retaining only the top map (Langmead & Salzberg 2012), from
298 this we recorded the number of mapped reads per kilobase of transcript per million
299 mapped reads. We considered viruses to show strong evidence of replication in the
300 host if they had at least 50 mapping siRNA reads with a size distribution sharply
301 peaked at 22nt (viRNAs are generated from replicating viruses by Dcr2). Following
302 Fauquet and Stanley (Fauquet & Stanley 2005), we defined contigs exhibiting less
303 than 90% nucleotide identity as separate viruses and those exhibiting greater than or
304 equal to 90% identity as strains of known viruses.
305
306 **PCR Validation and Testing**
307 A subset of contigs were chosen for manual validation. All chosen contigs met both of
308 the following conditions: the presence of mapping reads in the bumblebee small
309 RNAs (for the *B. terrestris*, *B. pascuorum* and *B. lucorum* pools; not a condition for
310 the mixed *Bombus* pool) or the transcriptomic RNAs (for the other *Bombus* pool
311 where small RNAs were not generated), and the closest blast match being viral RNA-
312 dependent RNA-polymerase. Internal primers for these contigs were generated using
313 primer3 (Untergasser et al. 2012) and amplification of the target was verified via
314 Sanger sequencing. See Supplementary Table 3 for PCR conditions and primer
315 sequences. Mayfield virus 1 and 2 were Sanger sequence validated over the entirety

316 of the contig. The Loch Morlich and River Liunaeg virus sequences were generated
317 by the connection of several disjoint contigs by Sanger sequencing. Black Hill virus
318 was excluded from further analysis as it was found that that the PCR reaction
319 amplified a host sequence that could not be visually differentiated from the virus
320 product.
321
322 **Phylogenetic Inference**
323 Following Cameron et al. (2007), we inferred the bumblebee phylogeny using
324 cytochrome oxidase I, elongation factor 1-alpha, opsin, phosphoenolpyruvate
325 carboxykinase, 16S and arginine kinase genes. To break up long branches and allow
326 dating, additional species not sampled in the field were added (see Supplementary
327 Table 4 for genbank accession numbers and species included). The DNA sequences
328 were aligned with MAFFT using the L-INS-i setting (Katoh et al. 2017; Katoh et al.
329 2005). The 6 gene alignments were then used to generate the phylogeny in BEAST
330 v2.4.5 (Bouckaert et al. 2014), treating each file as a separate partition, using
331 bModelTest (Bouckaert & Drummond 2015) with the 'transitionTransversion split'
332 setting and a calibrated Yule tree prior (Heled & Drummond 2012). An uncorrelated
333 lognormal relaxed clock was fitted to each partition, with exponential(λ=1) priors
334 placed over the mean rate and the default gamma(α=0.5396, β=0.3819) priors being
335 placed over the standard deviation (Drummond et al. 2006). The bumblebees were
336 constrained to be monophyletic, with the honeybee, *Apis mellifera*, as an outgroup. A
337 gamma (α=74.85889, β=0.4366812) distributed divergence time prior was placed
338 over the tMRCA of the *Bombus* clade, with parameters optimised to match the 2.5th
339 and 97.5th percentiles of the posterior distribution of ages previously estimated by
340 Hines (2008). Four separate runs of the MCMC were performed for 100,000,000 steps

341 from random starting trees, with the first 50,000,000 steps being discarded as burn in.
342 Convergence of the posterior among runs was assessed in Tracer v1.7 (Rambaut et al.
343 2017). The posterior distribution was thinned to 1000 trees.
344
345 For the virus phylogeny, amino acid sequences were inferred based on the translated
346 ORFs for regions predicted to contain RdRp motifs using the GenomeNet MOTIF
347 search function (Kanehisa et al. 2002) against the Pfam database (Finn et al. 2014),
348 with an expectation cut-off of 0.00001. If a virus had no annotated motifs, the
349 canonical GDD RdRp amino acid motif (Kamer & Argos 1984) was identified
350 manually. Additional virus species (Supplementary Table 5) were added to the
351 phylogeny to anchor species with short generated contigs, and to break up long
352 branches. Given the long evolutionary distance between the viruses, PROMALS3D
353 (Pei et al. 2008) was used to align viral sequences. The alignments were trimmed to
354 the first conserved secondary structural element at both ends as predicted by
355 PROMALS3D with the 0.95 conservation metric. Two of the novel viruses (Agassiz
356 Rock virus and Cnoc Mor virus) were not included in this phylogeny because the
357 section of the RdRp gene required fell outside the available contig. Given that it is
358 unclear whether there was a universal common ancestor of all RNA viruses (Koonin
359 et al. 2015), we aligned the sequences and generated the phylogeny twice, with and
360 without the negative sense RNA viruses (Supplementary Table 5). The trees serve
361 purely to quantify expected variance (under a Brownian motion model of evolution)
362 between closely related viruses. The deep splits in the phylogeny are poorly resolved
363 with RdRp data (Zanotto et al. 1996), due to the fast evolutionary rates of RNA
364 viruses, the considerable time since divergence and permutations in the RdRp
365 sequence (Gorbalenya et al. 2002). However, this should not overly bias the

366 conclusions as beyond a certain evolutionary distance, the viruses would be expected
367 to become essentially uncorrelated when averaged across the posterior
368 (Supplementary Table 6 for realised correlations).
369
370 Phylogenetic models used the BLOSUM62 rate matrix (Henikoff & Henikoff 1992)
371 with gamma distributed rate variation using 4 gamma categories, an uncorrelated
372 lognormal relaxed clock (Drummond et al. 2006) and a Yule tree prior. A CTMC rate
373 reference prior (Ferreira & Suchard 2008) was placed over the clock mean and an
374 exponential(λ=1) prior was placed over the standard deviation. The alpha parameter
375 of the gamma distributed rate variation was given an exponential(λ=1) prior.
376 Absolute dating of viral trees is difficult due to the inconsistency in estimated ages
377 provided by estimated clock rates and known orthologous insertions between sister
378 host species (Holmes 2003), but is not essential for our analysis, which depends only
379 on relative branch lengths. Nevertheless, we chose to use orthologous insertions to
380 provide approximate dates for our tree. To account for the estimated ages of RNA
381 viral families (Katzourakis & Gifford 2010), we set a uniform lognormal prior with
382 an offset of 97 Mya, a mean of 500 Mya and a logged standard deviation of 0.5 on the
383 age of the root of the tree including the negative sense RNA viruses and a lognormal
384 prior with an offset of 76 Mya, a mean of 500 Mya and a logged standard deviation of
385 0.5 on the age of the tree excluding them. Two partitiviruses (Rosellinia necatrix
386 partitivirus 2 and Raphanus sativus cryptic virus 1) known to have a common ancestor
387 older than 10 Mya (Chiba et al. 2011) were included for dating purposes. We placed a
388 diffuse lognormal prior with an offset of 10 Mya, a mean of 30 Mya and a logged
389 standard deviation of 0.5, on the age of the MRCA of these species. Both models
390 were run over 10 separate chains for 50,000,000 generations on a cluster in BEAST

391 v1.8.4 (Drummond et al. 2012), with 25,000,000 generations being discarded as burn-
392 in. Convergence of the posterior was assessed in Tracer v1.7 (Rambaut et al. 2017).
393 The posterior distributions were combined and thinned to 1000 trees.
394
395 **Prevalence Estimation**
396 Maximum likelihood prevalence and 2-log-likelihood confidence intervals were
397 estimated for each host/virus combination with more than one pool using the code
398 from Webster et al. (2015). As the samples were small pooled groups of individuals,
399 such that a PCR 'positive' represents one or more infections, we modelled the
400 prevalence using a "pooled binomial" likelihood (Ebert et al. 2010; Gibbs & Gower
401 1960; Thompson 1962). This approach requires that the underlying prevalence of a
402 virus is the same in all pools, which is unlikely for bumblebees sampled from
403 different locations. Estimates should therefore be treated with caution.
404
405 **Co-phylogenetic Mixed Model Analysis**
406 To test for the evolutionary effects on association, the presence/absence data and the
407 phylogenetic trees were analysed using a co-phylogenetic mixed model (Hadfield et
408 al. 2014) implemented in Stan (Carpenter et al. 2017). Our model is explicitly focused
409 at the individual-level, and the model's predictions represent the predicted probability
410 of infection within an individual of the species. This is in contrast with Hadfield et
411 al.'s original implementation where the focus was at the species- or population- level
412 and the model was estimating the probability that the parasite would be found in the
413 species or population at all. In all cases, the presented models showed no divergences,
414 acceptable Rhat and E-BFMI values and effective sample sizes of over 200.
415

416 We fitted host and virus phylogenetic effects, which measure the extent to which
417 variation in prevalence is clustered on the host and viral phylogenies respectively. We
418 also fitted host and viral evolutionary interaction effects, which measure the extent to
419 which related species have similar probabilities of infection in the sets of their
420 interaction partners. The final phylogenetic term fitted was a coevolutionary
421 interaction, which measures the extent to which related hosts are infected to similar
422 degrees by related viruses.
423
424 In addition to the phylogenetic terms, non-phylogenetic host and virus terms, an
425 interaction between these terms and a pool ID term were fitted. The non-phylogenetic
426 host and virus terms measure variation that can be partitioned between host species
427 and virus species in average infection risk that is not consistent with trait evolution by
428 Brownian motion. The interaction term measures variation that can be partitioned
429 between pairwise interactions between individual hosts and viruses that is not
430 consistent with the linear sum of their individual means from the non-phylogenetic
431 host and virus terms. The pool ID effect measures variation between pools in
432 infection risk averaged over all the viruses tested. As the pools combined hosts by
433 species rather than by location, so that some had individuals from multiple locations,
434 we treated each location and each realised combination of locations as levels of a
435 random effect, terming this the "spatial composition effect". This describes the
436 variation in average infection level between realised combinations of locations
437 averaged across viruses. Model 1 included all the viruses, Model 2 excluded the
438 negative sense RNA viruses and Model 3 fitted a pseudo-taxonomic model. In Model
439 3, the relationship among the viruses was represented by a polytomic viral tree with
440 arbitrary branch lengths (with a root-to tip distance of 1 unit, and equal length

441 between each taxonomic level) with the viruses being split first by their genomic type
442 (+ve sense RNA, -ve sense RNA and dsRNA) implying a covariance of 0 between
443 genome structures, followed by splitting by the putative viral clades identified by Shi
444 et al. (2016). This was done to test for potential bias caused the by the possibility of
445 systematic misidentification of the correct relationship between families in the
446 estimated viral trees.
447
448 The form of the models is shown below, where i indexes the data points, group_i
449 represents the level of a categorical variable that the ith pool belongs to, y_i represents
450 the 1/0 indicator for the presence or absence of infection in the ith pool, k_i represents
451 the number of individuals in the ith pool, p_i is the unmeasured probability of infection
452 of a single individual in the ith pool, y'_i is the estimated value of $\log_e(p_i/(1 - p_i))$, μ is
453 the global mean of the latent variable, ε is a normally distributed error term. All terms
454 were fitted as random effects (i.e. estimated by partial pooling). As above, a "pooled
455 binomial" likelihood was used (Ebert et al. 2010; Gibbs & Gower 1960; Thompson
456 1962).
457

$$y_i \sim \text{Bernoulli}(1 - (1 - p_i)^{k_i})$$

458 $$p_i = \exp(y'_i)/\exp(1 + y'_i)$$

459 $$y'_i = \mu + \text{host}_i + \text{virus}_i + \text{interaction}_i + \text{host phylogenetic effect}_i + \text{virus}$$
460 $$\text{phylogenetic effect}_i + \text{host evolutionary interaction effect}_i + \text{virus evolutionary}$$
461 $$\text{interaction}_i + \text{coevolutionary interaction}_i + \text{pool ID}_i + \text{species composition}_i + \varepsilon$$
462
463 All variance-covariance matrices were generated as described in Hadfield et al.
464 (2014), with the variance-covariance matrices scaled to correlation matrices. A

standard logistic prior was placed over the global intercept on the latent scale, μ, representing a flat prior on the probability scale. An exponential(λ=1) prior was placed on each variance term in the model. In the full model with all variances being estimated, this is equivalent to a gamma(α=11, β=1) prior over the total variance, which gives a prior mean variance of 11, and an appropriate prior on the standard deviation of a variable on the logit scale. Intraclass correlations, which represent the proportion of the variance explained by each effect, were calculated on the link scale (with an addition of $\pi^2/3$ to the denominator to account for the variance of the logistic distribution of the latent variable) from the model outputs and reported. Highest posterior density intervals were calculated by the SPIn method (Liu et al. 2015) and 90% credible intervals are reported as these are more robust to sampling in the tails of the posterior distribution (Stan Development Team 2017).

The total phylogenetic variance was calculated as:

$$(\sigma^2_{\text{host phylogenetic}} + \sigma^2_{\text{host interaction}} + \sigma^2_{\text{virus phylogenetic}} + \sigma^2_{\text{virus interaction}} + \sigma^2_{\text{coevolutionary interaction}})/(\sigma^2_{\text{total}} + \pi^2/3)$$

The total non-phylogenetic variance was calculated as:

$$(\sigma^2_{\text{host}} + \sigma^2_{\text{virus}} + \sigma^2_{\text{interaction}} + \sigma^2_{\text{poolID}} + \sigma^2_{\text{spatial composition}})/(\sigma_{\text{total}} + \pi^2/3)$$

Uncertainty in the inferred phylogenies was accounted for by direct marginalisation. This dramatically increased the runtime of the model, as, given the input of *H* host phylogenies and *V* viral phylogenies from their posterior distributions, the likelihood

of each datapoint has to be calculated *HV* times. As such, we included only 10 trees from each posterior, as a trade-off between runtime and accounting for uncertainty in the tree hypotheses. The marginalisation is show below, with **y** being the total vector of presences and absences, *H* being the number of host phylogenies used, *V* being the number viral phylogenies used, θ being all the non-variance-covariance parameters in the model, Ω_j being the set of variance-covariance matrices generated by the *j*th combination of host and virus phylogenies, Ω_{HV} representing the set of all variance-covariance matrices being marginalised over and $\mathfrak{L}$ representing a likelihood.

$$\mathfrak{L}(\boldsymbol{y} \mid \theta, \Omega_{HV}) = \sum_{j=1}^{HV} \frac{1}{HV} \mathfrak{L}(\boldsymbol{y} \mid \theta, \Omega_j)$$

Tongue Length Analysis

After finding that the posterior for the host evolutionary interaction was well resolved from zero, we designed a post-hoc test to attempt to detect signal for one of the obvious mechanistic explanations for this; structured transmission networks driven by evolutionarily conserved anatomical factors. We tested for an association between the tongue length differences between bumblebee species and the differences in their viral community structures, as a proxy for signal of differential transmission at flowers driven by evolutionarily conserved flower choice. Average tongue lengths for each bumblebee species except *Bombus bohemicus* and *Bombus cryptarum* were taken from Goulson, Hanley and Darvill (2005). No published tongue length could be found for *Bombus cryptarum*, so we assumed that it was identical to that of *Bombus lucorum*, a species of which it is near indistinguishable in the field. *Bombus bohemicus* was excluded from this analysis, because it is an inquiline parasite, and

514 therefore its ecology differs from the other species in such a way that tongue length
515 would not expected be expected to be correlated with the viral community distance.
516
517 In order to test for a correlation between tongue length and virus similarity, estimates
518 of the distance in viral communities between host species are required. These were
519 generated as follows: For each host-virus combination, the package 'prevalence' was
520 used to generate posterior draws of the underlying prevalences under a Beta (1,1)
521 prior. Then 1000 draws per species were taken from these sets of MCMC draws to
522 generate 1000 matrices of host-virus prevalences consistent with the raw data. For
523 each of these matrices, the distance between each species' viral community was
524 calculated by taking the vector of estimated prevalences for the 16 viruses of a given
525 species as a coordinate in a 16-dimensional space then calculating the Euclidean
526 distance between these points. The rank correlation (Kendall's τ-b) between each pair
527 of species' viral community distances and their tongue length distances was then
528 calculated, using the mantel function in the R package vegan. The point estimate
529 presented is the median of the 1000 initial correlations accounting for the uncertainty
530 in the underlying prevalences. The 95% confidence interval is the 2.5^{th} and 97.5^{th}
531 percentiles distribution of estimated correlations.
532
533 **Results**
534
535 RNA was extracted from 13 species of bumblebee from nine sites, to identify new
536 viruses, assay their prevalence and their pattern of distribution across host species and
537 to test whether the evolutionary histories of the viruses and hosts have impacted the
538 current distribution.

539

540 **Read and Assembly Statistics**

541 A total of 134,026,056 sequencing read pairs were generated for *Bombus lucorum*,
542 135,590,922 for *Bombus terrestris*, 128,670,194 for *Bombus pascuorum* and
543 26,838,390 for the other *Bombus* species with 0.37, 0.38, 3.36 and 15.12 percent of
544 reads mapping to the known viruses or the novel bee viruses found in the study. The
545 poly-A and DSN normalized datasets were unexpectedly highly correlated, given their
546 expected biases (1.000 for *Bombus terrestris*, 0.999 for *Bombus pascuorum* and 0.998
547 for *Bombus lucorum*) implying that the sequences results were highly consistent
548 irrespective of the selection method used.

549

550 **Previously Described Viruses Present in the Metagenomic Pools**

551 RNAseq reads mapped to three previously described bee viruses. The majority of
552 these reads mapped either to the Acute bee paralysis virus/Kashmir bee virus complex
553 (henceforth ABPV) (Bailey et al. 1963) or to Slow bee paralysis virus (SBPV) (Bailey
554 & Woods 1974). Additionally, in the mixed *Bombus* pool, reads were found mapping
555 to Hubei partiti-like virus 34 (HPLV34) a virus initially detected, though not named,
556 in honeybees by Cornman et al. (2012), then subsequently also reported in a sample
557 from Chinese landsnails by Shi et al (2016).

558

559 No RNAseq reads were mapped to Deformed wing virus – type A (Bailey & Ball
560 1991), Chronic bee paralysis virus (Bailey et al. 1963), Bee macula-like virus (de
561 Miranda et al. 2015), Ganda bee virus (Schoonvaere et al. 2016), Scaldis River bee
562 virus (Schoonvaere et al. 2016), Black queen cell virus (Bailey & Woods 1977), Apis
563 rhabdovirus 1 (Remnant et al. 2017), Apis rhabdovirus 2 (Remnant et al. 2017), Apis

564 bunyavirus 1 (Remnant et al. 2017), Apis bunyavirus 2 (Remnant et al. 2017), Apis
565 flavivirus (Remnant et al. 2017), Apis dicistrovirus (Remnant et al. 2017), Apis Nora
566 virus (Remnant et al. 2017) and members of the Lake Sinai virus complex (Runckel et
567 al. 2011). A small number of small RNA reads did map to these viruses, however, this
568 likely represents cross-mapping, given the lower stringency of 22nt reads. Two of the
569 viral contigs generated by the *de novo* assembly had high similarity to previously
570 described plant viruses; both RNAs of White clover cryptic virus 2 (Boccardo et al.
571 1985) (96% identity), both RNAs of strain of Arabis mosaic virus
572 (MH614320/MH614321) (Smith & Markham 1944) distant to previously sequenced
573 strains (91% identity) and a strain of Red Clover nepovirus A (MH614312) (Koloniuk
574 et al. 2018) distant to previously sequenced strains (90% identity).
575
576 **Putative Novel Viral-like Sequences**
577 We identified 37 putative novel viral contigs, four mapping to DNA viruses (4
578 densovirus-like contigs) and 33 to RNA viruses (4 Reo group contigs, 2 Toti-Chryso
579 group contigs, 4 Bunya-Arena group contigs, 1 Orthomyxoviridae-like contig, 8
580 Hepe-Virga group contigs, 12 Picorna-Calici group contigs and 2 Tombus-Noda
581 group contigs). Based on the supposition that a contig represents a separate virus if it
582 maps to a different viral grouping than the other contigs, or if it can be aligned to all
583 other contigs within its assigned viral grouping, this represents 30 new viruses with
584 seven remaining contigs that may represent other genomic regions of these 30 viruses
585 or separate viruses that cannot be confirmed as such. See Table 1 for information on
586 the viruses tested for prevalence using PCR and Supplementary Table 7 for detailed
587 information on all of the identified contigs. The numbers of reads mapping these
588 contigs were variable and are shown in Table 2.

589

590 **Table 1** The names, genome structures and groupings (following Shi et al. (2016)) of the newly

591 discovered viruses for which prevalence was assessed.

Putative viral contig	Abbreviations	Clade	Genome structure
Agassiz Rock virus	ARV	Reo	dsRNA
Elf Loch virus	ELV	Reo	dsRNA
Dumyat virus	DV	Toti-Chryso	dsRNA
Sheriffmuir virus	SV	Toti-Chryso	dsRNA
Clamshell Cave virus	CCV	Bunya-Arena	- ssRNA
Allermuir Hill virus 1	AHV1	Hepe-Virga	+ssRNA
Allermuir Hill virus 2	AHV2	Hepe-Virga	+ssRNA
Allermuir Hill virus 3	AHV3	Hepe-Virga	+ssRNA
Mill Lade virus	MLV	Hepe-Virga	+ssRNA
Boghill Burn virus	BBV	Picorna-Calici	+ssRNA
Gorebridge virus	GV	Picorna-Calici	+ssRNA
Loch Morlich virus	LMV	Picorna-Calici	+ssRNA
Mayfield virus 1	MV1	Picorna-Calici	+ssRNA
Mayfield virus 2	MV2	Picorna-Calici	+ssRNA
River Liunaeg virus	RLV	Picorna-Calici	+ssRNA
Castleton Burn virus	CBV	Tombus-Noda	+ssRNA

592

593 **Table 2** The RNAseq reads per kilobase per mapped million reads in the *Bombus terrestris*, *Bombus*

594 *lucorum, Bombus* pascuorum and mixed *Bombus* pools. Structural zeros are indicated by dashes, zeros

595 in the table indicate below 0.005. Contigs with names in bold meet the criterion of having at least 50

596 mapping small RNA reads with a sharp peak in the size distribution at 22nt in *Bombus terrestris,*

597 *Bombus lucorum* and *Bombus pascuorum* providing evidence of replication (see main text).

Putative viral contig	Accession number	*Bombus terrestris*	*Bombus lucorum*	*Bombus pascuorum*	mixed *Bombus*
Bombus-associated Densovirus-like Contig 1	MH614322	0.09	-	-	12.34
Bombus-associated Densovirus-like Contig 2	MH614323	0.39	2.66	-	14.23
Agassiz Rock virus	MH614287	3.77	0.49	-	-
Cnoc Mor virus	MH614297	0.97	-	-	20.38
Bombus-associated Reoviridae-like Contig 1	MH614298	0.42	0.03	-	1.63
Elf Loch virus	MH614300	-	0.00	1.00	0.05
Dumyat virus	MH614299	-	-	-	10.30
Sheriffmuir virus	MH614317	-	-	-	2.55

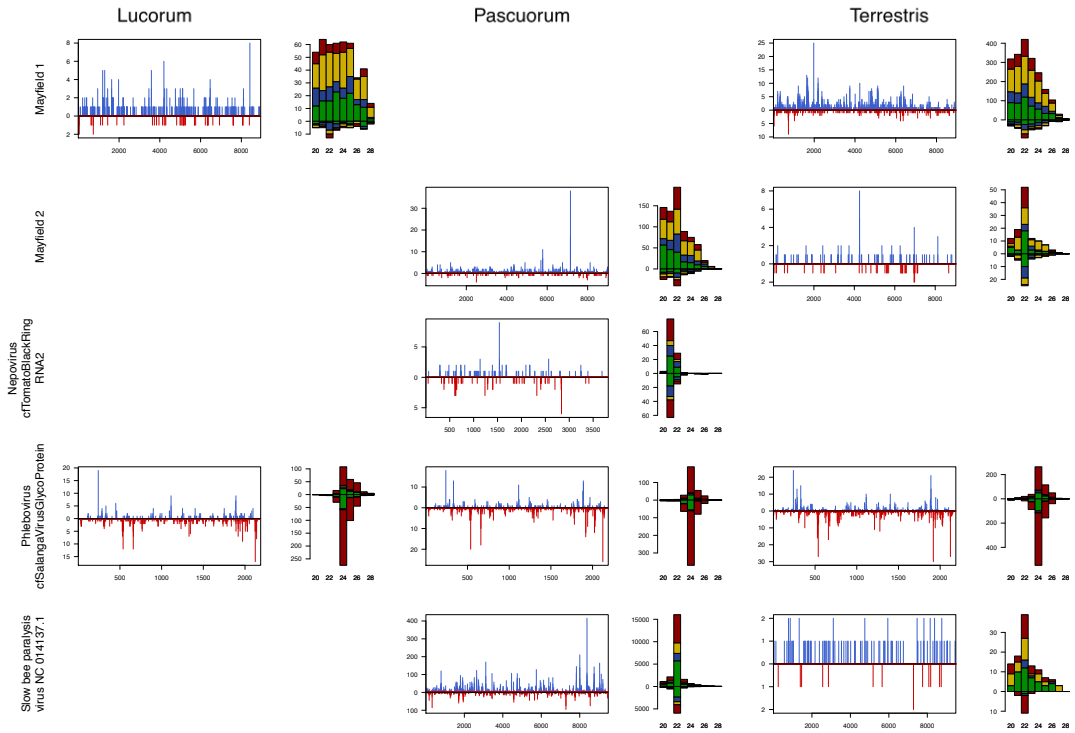
Clamshell Cave virus	MH614294	0.07	-	-	2.10
Bombus-associated Bunyaviridae-like Contig 1	MH614295	0.70	-	-	5.61
Bombus-associated Bunyaviridae-like Contig 2	MH614296	-	-	-	12.13
Bombus-associated Phlebovirus-like Contig 1	MH614315	3.07	2.77	0.95	1.63
Bombus-associated Orthomyxovirus-like Contig 1	MH614314	-	-	0.44	-
Allermuir Hill virus 1	MH614288	15.27	0.64	0.02	1.50
Allermuir Hill virus 2	MH614289	0.01	0.03	12.71	0.13
Allermuir Hill virus 3	MH614290	0.40	2.62	0.50	3.35
Mill Lade virus	MH614306	0.40	0.48	0.03	7.28
Bombus-associated Virga-like Contig 1	MH614308	-	0.65	-	-
Bombus-associated Virga-like Contig 2	MH614309	-	0.33	-	-
Bombus-associated Virga-like Contig 3	MH614318	16.03	4.83	0.52	90.37
Bombus-associated Virga-like Contig 4	MH614319	1.66	0.11	-	0.54
Black Hill virus	MH614291	-	-	-	2.96
Boghill Burn virus	MH614292	0.00	10.92	0.00	0.11
Gorebridge virus	MH614301	2.97	0.02	-	0.15
Bombus-associated Picornavirus-like Contig 1	MH614302	6.27	0.02	-	0.29
Loch Morlich virus	MH614303	0.00	-	-	7.83
Mayfield virus 1	MH614304	391.31	232.93	0.59	7.67
Mayfield virus 2	MH614305	4.87	0.79	336.97	558.82
Bombus-associated Nepovirus-like Contig 1	MH614310	1.72	1.02	0.09	1.70
Bombus-associated Nepovirus-like Contig 2	MH614311	0.33	0.39	0.11	0.21
Bombus-associated Picornavirus-like Contig 2	MH614316	0.11	0.29	1.02	0.04
River Liunaeg virus	MH614307	0.24	0.47	0.01	9.58
Castleton Burn virus	MH614293	2.39	12.30	8.10	122.99
Bombus-associated Nodavirus-like Contig 1	MH614313	-	-	-	1.50

siRNA-based Evidence for Infection

RNA interference is an important component of antiviral defence in arthropods (Bronkhorst et al. 2012). As part of this defence mechanism, homologs of *Drosophila* Dicer-2 cleave dsRNA, usually in the form of replication intermediates, giving rise to a characteristically narrow and sharply peaked distribution of virus-derived small RNAs. Thus the presence of such small RNAs from both strands of an ssRNA virus provide compelling evidence that the virus was replicating. In bumblebees the

606 characteristic Dicer-mediated viral siRNAs peak sharply at 22nt (Remnant et al.
607 2017), and the viruses that displayed at least 50 characteristic viral siRNAs are
608 marked in Table 2. The distribution of the mapped small RNA reads is shown in
609 Figure 3 for all viruses where the siRNAs are described in the main text, with full data
610 in Supplementary Figure 1.
611
612 In the three bumblebee species with siRNA data, a sequence similar to a phlebovirus
613 glycoprotein (AEL29653.1) displayed >50 siRNA reads. However, the size spectra of
614 these reads is centered on 24nt with a strong bias for a 5' terminal uracil, with the
615 antisense mapping orientation being more prevalent. This 5' U-bias is consistent with
616 insect piRNAs (Brennecke et al. 2007), and the predominant antisense orientation is
617 consistent with the piRNA mapping pattern to endogenous viral elements (EVE) in
618 mosquitoes (Suzuki et al. 2017). However, the size of piRNAs in bumblebees is
619 generally larger than this (Lewis et al. 2018). This sequence is therefore potentially an
620 EVE that has either been gained multiple times or has been maintained in the
621 bumblebee genome since at least the *B. pascuorum*-*B. terrestris*/*B. lucorum* split.
622
623 It is notable that the size distribution of viral siRNAs is less sharply peaked in
624 Mayfield virus 1, Mayfield virus 2 and Slow bee paralysis virus (excepting Mayfield
625 virus 1 in *B. lucorum*, which is sharply peaked), with broad 'shoulders'. This is
626 reminiscent of the pattern seen for Drosophila C virus and Drosophila Nora virus in
627 wild-collected D. *melanogaster* (Webster et al. 2015), both of which contain a viral
628 suppressor of RNAi (van Rij et al. 2006; van Mierlo et al. 2012).
629

630 *B. pascuorum* also had siRNA reads mapping to a sequence with 19% identity to
631 Tomato black ring virus (CAA56792.1). However, the read length spectra were
632 sharply peaked at 21nt, rather than the 22nt of bumblebee viRNAs. This is consistent
633 with siRNA's produced from DLC4, the key antiviral dicer in *Arabidopsis thaliana*
634 (Deleris et al. 2006) implying acquisition of the small RNAs through nectar or pollen
635 contamination.



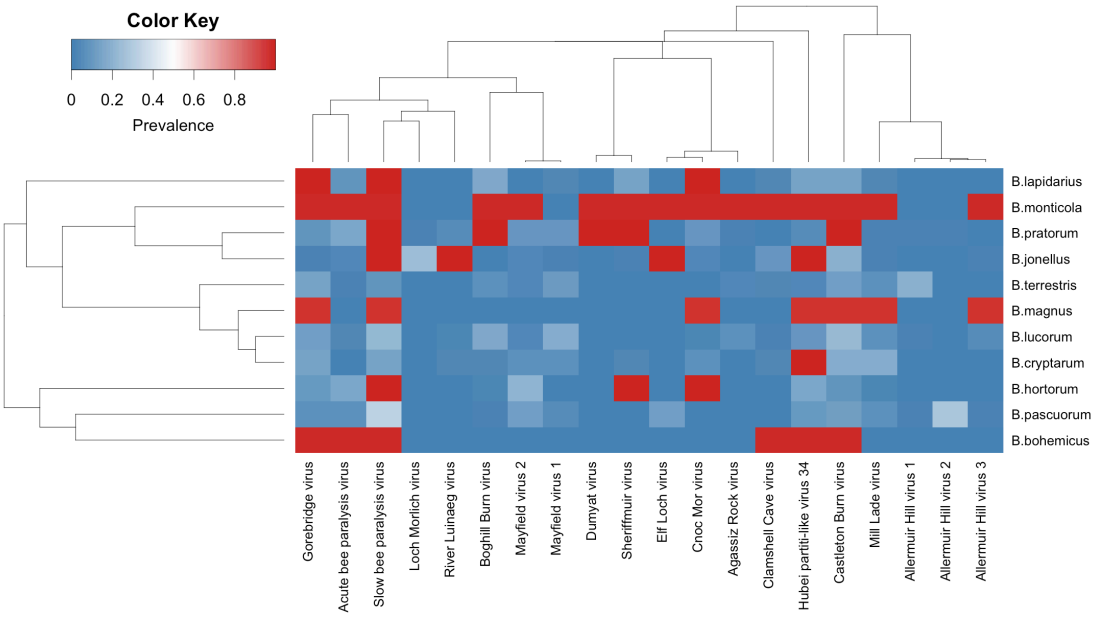
636

637 **Figure 2** The mapping of small RNA reads to Mayfield virus 1, Mayfield virus 2, a contig similar to
638 Tomato black ring virus, a contig similar to a phlebovirus glycoprotein and Slow bee paralysis virus.
639 Blue lines represent reads mapping to the positive-sense strand at that genomic position, red lines
640 represent reads mapping to the negative sense strand. The histogram of read size spectra shows the
641 count of reads of each length mapping in the positive (above) and negative (below) directions. The
642 colouring of each bar shows the counts of the reads beginning with each 5' base (red-U, blue-C, green-
643 A, yellow-G).

644

645 **Prevalence**

646 Species level prevalences differed dramatically among the different viruses (Figure
647 3). Prevalences were generally low to intermediate, with modal viral prevalences for
648 most host-virus combinations being below 15%. Slow bee paralysis virus was by far
649 the most common virus in the sample, with estimated prevalences of greater than 25%
650 in multiple species. Our ability to estimate the prevalence of common viruses is
651 limited by the pooling, leading us to only be able to assign lower bounds to
652 prevalences in these cases, but in 7 of 11 species, all pools were positive for SBPV.
653 Acute bee paralysis virus, Hubei partiti-like virus 34, Castleton Burn virus,
654 Gorebridge virus, Mayfield virus 1 and Mayfield virus 2 all reached 15-25%
655 prevalences in multiple species. Several viruses showed strong signals of species
656 specificity, having very low to zero prevalences in multiple host species but high
657 prevalences in others. Examples of this pattern include Allermuir Hill virus 1 in *B.*
658 *terrestris,* Allermuir Hill virus 2 in *B. pascuorum*, Allermuir Hill virus 3 in *B. magnus*
659 and *B. monticola*, as well as Loch Morlich virus and River Luinaeg virus in *B.*
660 *jonellus*.
661



662

663 **Figure 3** A heatmap of maximum likelihood estimates for prevalence. Hosts and viruses are ordered by

664 phylogenetic relatedness, the trees represent the maximum clade credibility topology. Squares in red
665 with maximum likelihood estimates of the prevalence of 1 correspond to cases where all pools were
666 positive. The maximum likelihood estimate is likely extremely upwardly biased in this case.
667

668 **Host-Pathogen Co-phylogenetic Models**

669 All models that included a virus phylogeny term gave qualitatively similar results
670 (Figure 4, Table 3). This suggests that the results are robust to both phylogenetic
671 uncertainty and the assumption of a common ancestor of all RNA viruses. For this
672 reason, for the rest of this section, estimates will be given from the model containing
673 the estimated phylogeny with all the RNA viruses included. All estimates represent
674 the percentage of the total variance in the model (the sum of all estimated variance
675 components adjusted for the variance of the link function by the addition of $\pi^2/3$)
676 explained by a term. In all cases, the presented point estimate is the posterior mean,
677 and 90% shortest posterior density intervals (Liu et al. 2015) are presented following
678 in square brackets. Shortest posterior density intervals are a variant of highest
679 posterior density intervals and describe the shortest possible interval containing (in
680 this case) 90% of the probability density for the parameter. We present 90% intervals
681 rather than the standard 95% intervals, as 95% intervals calculated from simulation
682 draws are less computationally stable (Stan Development Team 2017). In all cases but
683 the virus phylogenetic effect, the posterior estimates for the proportion of variance
684 explained by each effect differed strongly from their induced priors (Supplementary
685 Figure 2).
686

687 **Summary of Model Results**

688 We find evidence that which viral species infects a host, the specific interaction
689 between individual hosts and individual viruses and related hosts having similar

690 prevalences with the same sets of viruses all explain variation in infection prevalence.

691

692 **Total Evolutionarily-associated Variation**

693 In the models containing the virus phylogeny, approximately a quarter (25.9% [11.6-
694 40.4]) of the total variation in prevalence was explained by terms accounting for the
695 evolutionary histories of hosts and viruses (host phylogenetic effect, virus
696 phylogenetic effect, host evolutionary interaction effect, virus evolutionary interaction
697 effect and coevolutionary interaction).

698

699 **Host and Virus Level Effects**

700 The host and virus phylogenetic effects measure the extent to which related hosts
701 have similar average prevalences of virus infection and related viruses have similar
702 average prevalences across hosts. The host and virus non-phylogenetic terms measure
703 the extent to which hosts and viruses differ in their average infection levels in manner
704 not consistent with evolution by Brownian motion along a phylogeny. The host
705 species and phylogenetic effects explained a small proportion of the total variance in
706 infection probability (species: 1.9%, [0.0-4.6]; phylogenetic: 2.9%, [0.0-6.5]). The
707 shape of the posterior distributions for the two parameters visualised in Figure 6,
708 makes it clear that the most credible values for both of these parameters are 0. While
709 it is unlikely that there is no variation in average prevalence between host species, it is
710 clear that the amount of prevalence explained by hosts differing in their infection
711 levels averaged across viruses is small relative to the other effects.

712

713 The virus species and phylogenetic effects explained a larger proportion of the total
714 variance in infection probability, with non-phylogenetic variation dominating, but

were imprecisely estimated (species: 13.8%, [3.8-23.7]; phylogenetic: 8.5%, [0.0-17.8]). The posterior density for the phylogenetic effect is concentrated at 0. So, while it is clear there is a virus species effect, the data does not appear informative for the presence or absence of a viral phylogenetic effect in this system. The posterior draws for the viral species and phylogenetic effect were negatively correlated within iterations, indicating that the model had difficulty partitioning the two. This partial non-identifiablility explains the broad posteriors on both.

Interaction Effects

A host evolutionary interaction effect measures the extent to which more closely related hosts have more similar prevalences with the same sets of viruses, and a virus evolutionary interaction effect measures the extent to which more closely related viruses infect the same sets of hosts to more similar degrees. A coevolutionary interaction term measures the extent to which more related hosts are infected to similar degrees with viruses that are themselves related. The non-phylogenetic interaction term measures the extent to which there is variance in the mean prevalence of specific host-virus pairings, that are is not consistent with the other interaction effects.

There was little evidence for a virus evolutionary interaction effect or coevolutionary interaction having a large effect on the observed prevalences (virus: 2.9% [0.0-6.1]; coevolutionary: 2.0% [0.0-4.6]). In both cases, the marginal posterior distributions were peaked at 0.

There was evidence for a host evolutionary interaction explaining some of the total

740 variance in prevalence (9.5% [0.0-18.3]). This is the only parameter in the model
741 where the estimated size of the effect depended strongly on the specific treatment of
742 the virus phylogeny (see Figure 4), and indeed whether the lower 90% bound of the
743 credible interval rounded to 0.0 or 0.1 depended on the phylogenetic matrix (or set of
744 phylogenetic matrices) inputted. The marginal posterior distribution of the parameter
745 was concentrated at lower values when the estimated phylogeny including the
746 negative-sense RNA viruses was used, and at higher values in the other two cases.
747 However, irrespective of the choice of virus phylogeny, the mode of the distribution
748 and majority of the density was distant from zero, implying that the effect is likely to
749 be biologically relevant. As the virus phylogenies themselves are not actually directly
750 involved in this term, this must be due to the partitioning of variance across other
751 terms being cryptically different depending on the assumptions about the virus
752 phylogeny.
753
754 There was also a clear non-phylogenetic interaction (15.8% [5.8-25.4]), implying that
755 much of the variation in prevalence is due to the specifics of individual host-virus
756 combinations.
757
758 As with the virus species effect and the virus phylogenetic effect, the MCMC draws
759 for the proportion of variance explained by the host evolutionary interaction and non-
760 phylogenetic interaction were negatively correlated within an iteration, implying that
761 separating these parameters was proving difficult. While this lead to a diffuse
762 posterior with wide credible intervals for both, they remain individually interpretable,
763 and both effects appear present simultaneously.

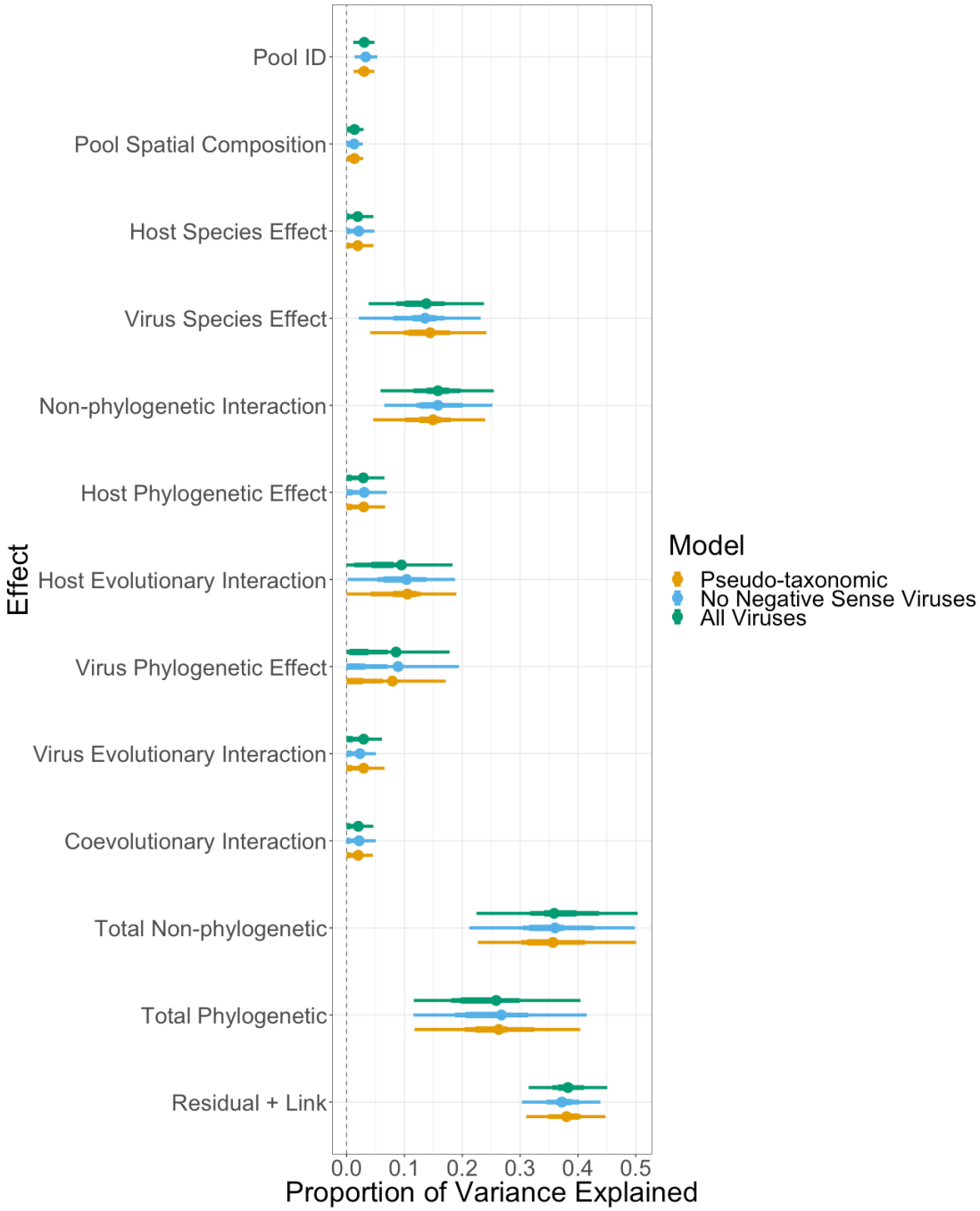


Figure 4 Comparison of estimated proportion of variance in prevalence explained by different parameters between models. As each factor explains a proportion of the attributed variance in the model total over all factors must sum to 1. For each parameter, the circle represents the modal estimate, the thick bars represent the 50% shortest posterior density interval and the thin bars represent the 90% shortest posterior density interval. "Pool ID" is the proportion of the total variation in prevalence explained by pools within species differing in the degree to which they were infected by viruses. "Spatial Composition" is the proportion of the total variation in prevalence explained by the combination of locations from which the bees in the pool originate. "Host Species Effect" is the

773 proportion of variation in prevalence explained by hosts having different average viral prevalences.
774 "Virus Species Effect" is the proportion of variation in prevalence explained by viruses differing in
775 their average prevalences. "Non-phylogenetic Interaction" is the proportion of variation in prevalence
776 explained by host-virus combinations differing in the their average prevalences beyond that which
777 would be expected by their host and virus species effects alone. "Host Phylogenetic Effect" is the
778 proportion of variation in prevalence explained by hosts having average viral prevalences correlated
779 across the host phylogeny. "Virus Phylogenetic Effect" is the proportion of variation in prevalence
780 explained by viruses having average prevalences correlated across the viral phylogeny. "Host
781 Evolutionary Interaction" is the proportion of variation in prevalence explained by related hosts having
782 correlated viral assemblages. "Virus Evolutionary Interaction" is the proportion of the variation
783 explained by related viruses having correlated host assemblages. "Coevolutionary Interaction" is the
784 proportion of the variation explained by related hosts having similar prevalences of related viruses.
785 "Total Non-phylogenetic" is the proportion of the variation that can be explained by terms not
786 involving the host and virus phylogeny and excluding the residual ("Host Species Effect", "Virus
787 Species Effect", "Pool ID", "Spatial Composition Effect", "Non-phylogenetic Interaction"). "Total
788 phylogenetic" is the proportion of the variation that can be explained by terms involving a host or virus
789 phylogeny ("Host Phylogenetic Effect", "Virus Phylogenetic Effect", "Host Evolutionary Interaction",
790 "Virus Evolutionary Interaction", "Coevolutionary Interaction"). "Residual + Link" is the proportion of
791 the total variance that is explained by the residual variance and variance of the logistic distribution
792 ($\pi^2/3$).
793
794

795 **Table 3** Mean estimates for the intra-class correlations of each variance component. The point estimate
796 is the posterior mean, the numbers in brackets represent the 90% shortest posterior density interval.

	All Viruses	No Negative Sense Viruses	Pseudo-taxonomic
Virus Species Effect	13.8 (3.8, 23.7)	13.6 (2.1, 23.2)	14.4 (4.1, 24.2)
Virus Phylogenetic Effect	8.5 (0, 17.8)	8.9 (0, 19.4)	7.9 (0, 17.1)
Host Species Effect	1.9 (0, 4.6)	2.1 (0, 4.8)	1.9 (0, 4.6)
Host Phylogenetic Effect	2.9 (0, 6.5)	3.0 (0, 6.9)	2.9 (0, 6.6)
Virus Evolutionary Interaction	2.9 (0, 6.1)	2.3 (0, 5.1)	2.9 (0, 6.6)
Host Evolutionary Interaction	9.5 (0, 18.3)	10.4 (0.2, 18.8)	10.5 (0, 19.0)
Non-phylogenetic Interaction	15.8 (5.8, 25.4)	15.8 (6.5, 25.2)	14.9 (4.6, 24.0)
Coevolutionary Interaction	2.0 (0, 4.6)	2.2 (0, 5.0)	2.0 (0, 4.5)
Pool ID	3.0 (1.1, 4.8)	3.3 (1.4, 5.3)	3.0 (1.2, 4.8)
Pool Spatial Composition	1.4 (0, 2.9)	1.3 (0, 2.8)	1.3 (0, 2.9)
Residual + Link	38.3 (31.5, 45.1)	37.2 (30.3, 43.9)	38 (31, 44.7)
Total Non-phylogenetic	35.9 (22.4, 50.3)	36 (21.2, 49.8)	35.7 (22.7, 50.1)
Total Phylogenetic	25.9 (11.6, 40.4)	26.8 (11.6, 41.5)	26.3 (11.7, 40.4)

797

798

799 **Tongue Length-Viral Community Correlation**

800 Given that the co-phylogenetic model found that related hosts share viral
801 communities, one potential mechanism for this is phylogenetically-biased exposure,
802 driven by phylogenetically correlated floral preferences. If bumblebee species with
803 similar flower preferences had similar viral communities, it would be expected that
804 there would be a positive correlation between tongue length similarity (as this is an
805 important factor in floral preference) and viral community similarity between pairs of
806 species. The point estimate of the correlation between the two distances was small
807 and negative (-0.06), but the 95% confidence intervals for that point estimate
808 overlapped zero (-0.13, 0.00), so given the uncertainty in the data, a correlation of
809 zero cannot be rejected. None-the-less, given this data, a strong positive relationship
810 between tongue length and viral community similarity seems unlikely, a result

811 inconsistent with phylogenetically-biased exposure driven by tongue length-mediated
812 floral choice.
813
814 **Discussion**
815 Using wild bumblebee species that share transmission opportunities, we have shown
816 that variation in the prevalence of infection in the wild is explained by related hosts
817 being infected with the same viruses to similar degrees, viruses differing in their
818 average prevalence and individual virus-host pairings having greater or lesser
819 prevalence than would be expected by the interaction of the host and virus species
820 effects alone.
821
822 **Virus Discovery**
823 There is now an extensive diversity of viruses known in bees, with most new studies
824 finding novel viruses (Cornman et al. 2012; Mordecai et al. 2015; Remnant et al.
825 2017; Runckel et al. 2011; Schoonvaere et al. 2016; Schoonvaere et al. 2018; Roberts
826 et al. 2018). We have found 37 novel putative viral contigs in the transcriptomes of
827 wild-caught bumblebees from across Scotland, suggesting that virus discovery in this
828 taxonomic group is far from saturation. As with any metagenomic study, it is hard to
829 be confident that the virus-like contigs represent real infections of the sampled host,
830 rather than surface or gut contaminants. However, the presence of 22nt virus siRNAs,
831 generated from double-stranded viruses by Dicer as part of an antiviral response in the
832 host, provides compelling evidence that at least 10 of these contigs (Densovirus 2 and
833 3, Elf Loch virus, Allermuir Hill virus 1, 2 and 3, Mill Lade virus, Mayfield virus 1
834 and 2, and Castleton Burn virus) represent active viral infections in bumblebees.
835

836 Mites and nematodes both parasitise bumblebees and therefore could potentially be an
837 alternative source of the small RNAs. Mite viRNAs are reported to be centered at
838 24nt (Remnant et al. 2017), and could therefore not produce the small RNA patterns
839 observed. Nematode viRNAs are centered at 22nt, like bumblebee viRNAs, (Félix et
840 al. 2011) and thus could potentially produce this pattern. While, outside of queens
841 infected with *Sphaerularia bombi*, nematode infection of wild bumblebees appears to
842 be very rare (Rao et al. 2017), nematodes cannot be categorically ruled out as a source
843 of the observed small RNAs. One contig's (MH614312) mapped small RNAs were
844 centered at 21nt, and the closest known virus was a nepovirus of plants. As DCL4, a
845 major plant Dicer, produces viRNAs of this size (Wang et al. 2011), this is consistent
846 with that particular virus being a plant virus, which was transferred in collected nectar
847 or pollen.
848
849 **Phylogenetic Effects**
850 We found no evidence for a large host phylogenetic effect, where related hosts have
851 correlated average viral prevalences. To the best of our knowledge no studies have
852 previously applied these methods to viruses sampled from wild animals. However,
853 other traits relating to viral disease in a series of studies in *Drosophila* species under
854 experimental conditions have consistently detected host phylogenetic effects in
855 factors that would be expected to be correlated with prevalence in the wild, such as
856 infection probability (Longdon et al. 2011), virulence and viral load (Longdon et al.
857 2015) and viral load alone (Roberts et al. 2018). However, two of these studies
858 focused on a single isolate of Drosophila C virus. Therefore, the variation that they
859 attribute to a phylogenetic effect may be partitioned into the host evolutionary
860 interaction in our study, as a host evolutionary interaction is equivalent to separate

861 inconsistent host phylogenetic effects for each interaction partner. Our were not
862 particularly informative for the presence or absence of a virus phylogenetic effect,
863 with the posterior being very diffuse with a majority of the density near zero. This
864 appears partly due to difficulties partitioning the variation between the virus species
865 effect and virus phylogenetic effect. Irrespective of the cause, we can make no strong
866 statements about whether related viruses exhibit similar prevalences across hosts from
867 this dataset.
868
869 **Host Species Effect**
870 There was little evidence for an important host species effect, implying that hosts do
871 not strongly vary in the average degree to which they are infected with viruses. The
872 previous studies using these methods have universally found host species effects
873 (Hadfield et al. 2014; Waxman et al. 2014). However, as both of these studies have
874 used mammal-eukaryotic parasite datasets, the degree of relevance for them as a
875 comparison in unclear. Experimental evidence from virus studies across drosophilid
876 flies have found weak to zero host species effects on the titre of sigma viruses
877 (Longdon et al. 2011) and Drosophila C virus virulence (Longdon et al. 2015) but
878 considerably larger host species effects on Drosophila C virus load (Longdon et al.
879 2015). This between-study variation potentially indicates a reason we did not detect a
880 host species effect. With a single pathogen, the average and particular degrees of
881 variation in infection between hosts are identical. As soon as multiple pathogens are
882 involved, they diverge, such that it is possible for there to be no variation in the
883 average prevalence between hosts, but still considerable variation in the prevalence of
884 particular viruses between hosts, which is consistent with the presence of a host

885 species effect in correlates of prevalence in some viruses but not others, as noted
886 above.
887
888 **Virus Species Effect**
889 A clear virus species effect was detected in the dataset, despite the uncertainty added
890 by the difficulty partitioning the virus species and phylogenetic effects. Therefore,
891 viruses differed in their prevalences averaged across hosts. This is not a surprising
892 result as viruses differ in host range (Bandín & Dopazo 2011), virulence (Langsjoen
893 et al. 2018; Baker & Antonovics 2012) and infectious period (Baker & Antonovics
894 2012) at both the species and the strain level. Variation in host range changes the size
895 of the host pool available for infection, and variation in virulence and infectious
896 period both change the length of time any infected host is available for sampling. All
897 these factors would be expected to drive consistent differences in long-run
898 prevalences between viruses. Additionally, as our sites were only sampled once,
899 short-term effects will also drive between virus variation. Any virus that was
900 experiencing an epizootic at the time of sampling will be overrepresented relative to
901 its long-run prevalence, further increasing the between virus variation.
902
903 **Host Evolutionary Interaction Effect**
904 A host assemblage effect was found, where phylogenetically related hosts share viral
905 assemblages, showing that more closely related hosts are more similar in virus
906 prevalence for groups of viruses. The statistical machinery required for estimating this
907 effect is quite new and, as such in the disease ecology field, has predominantly been
908 applied to mammal-parasite and plant-parasite systems where there are good datasets
909 already existing. None-the-less, host evolutionary interactions have always been

910 found when searched for using these methods (Waxman et al. 2014; Hadfield et al.
911 2014) and analogous effects are commonly found using different methods (Davies &
912 Pedersen 2008; Huang et al. 2014; Cooper et al. 2012). In a system where these
913 viruses were host limited, this pattern could be explained by preferential host shifting,
914 where parasites more frequently gain the ability to infect hosts closely related to ones
915 that they are already capable of infecting. Preferential host shifting is known to be a
916 general phenomenon, and has been observed in macroparasites, viruses and
917 protozoans (see Longdon et al. 2014 and the references within).
918
919 While some of the viruses in this study were not detected in a subset of host species,
920 most of the viruses found here appear to genuinely be multihost viruses, with the
921 majority being detected in over half the sampled species. Given this, a combination of
922 biased cross-species transmission and preferential host shifting appears a better
923 explanation in this system. Biased cross-species transmission occurs when
924 transmission occurs more frequently between some species that a pathogen is already
925 capable of infecting than others. This biased cross-species transmission could be
926 driven by two non-exclusive mechanisms: phylogenetically-biased transmission
927 probabilities and phylogenetically-biased exposure.
928
929 Phylogenetically-biased transmission probabilities occurs when cross-species
930 transmission is more frequent among close relatives, due to the probability of
931 infection after contact with the virus being similar between related species. Related
932 hosts present correlated environments from the perspective of the virus at the
933 molecular and anatomical level, therefore adaptation to one should provide
934 corresponding fitness increases on the other. Experimental results have shown that

935 correlated mutations occur on viral entry to related hosts (Longdon et al. 2018),
936 implying that this cross-adaptation does occur. However, this is probabilistic, and
937 different routes the mutations fixed on entry can differ between replicate entries
938 (Longdon et al. 2018; Streicker et al. 2012). Therefore, if there is antagonistic
939 pleiotropy between mutations that are adaptive in two different groups of hosts and
940 cross-adaptation predicts the probability of successful infection on contact, then a
941 phylogenetically-biased transmission network will result.
942
943 Phylogenetically-biased exposure represents an evolutionarily-driven ecological
944 phenomenon that biases cross-species transmission rates, mediated by niche overlap.
945 Contaminated flowers are likely to be an important source of intra- and inter- specific
946 pathogen transmission in bumblebees and pollinators more generally (Durrer &
947 Schmid-Hempel 1994; Graystock et al. 2015; McArt et al. 2014). The flower
948 visitation network has been shown to be associated with the partitioning of genetic
949 diversity of *Crithidia bombi* between bumblebee hosts (Salathé & Schmid-Hempel
950 2011), and the network itself is highly structured, though temporally variable (Ruiz-
951 González et al. 2012). Different bumblebee species show tongue length differences,
952 which are phylogenetically associated (Harmon-Threatt & Ackerly 2013), and the
953 differences in tongue length correlate with differential flower usage between
954 bumblebee species (Goulson et al. 2008; Inouye 1978). If infection occurs at
955 contaminated flowers, the structuring of the flower usage network could cause
956 different flowers to build up different surface viral communities. This could drive
957 consistent phylogenetically-correlated differences in viral infection rates through
958 differential exposure.
959

960 Post-hoc testing did not find a positive relationship between tongue length
961 dissimilarity (a rough proxy for species-level flower choice dissimilarity) and viral
962 community composition dissimilarity, which provides some evidence against
963 phylogenetically-biased exposure as the causative mechanism. However, the study
964 design in this case is not optimal for disentangling biased transmission probabilities
965 and biased exposure, as species were sampled from different locations at a single
966 timepoint and prevalence of the viruses varied spatially. Given this, drawing strong
967 conclusions as to the relative impact of the two mechanisms outlined above based on
968 this data would be premature. Similarly, the subgenus *Psithyrus* contains socially
969 parasitic species that are coevolved to parasitise particular social bumblebee species,
970 which could also lead to phylogenetically-correlated differences in viral infection
971 rates. We were unable to test whether socially parasitic cuckoo bumblebee species
972 have similar viral communities to their hosts, as our study included only a single
973 parasitic bumblebee species, *B. bohemicus*, but the possibility of brood parasitism
974 being an important driver of between-colony disease transmission is worth further
975 study.
976
977 **Virus Evolutionary Interaction Effect and Coevolutionary Interaction**
978 We found no evidence of a large virus evolutionary interaction or coevolutionary
979 interaction. This is largely unsurprising as it would appear implausible that the host
980 assemblages have been conserved over evolutionary time, as the deep splits in the
981 viral families predate the most recent common ancestor of bumblebees by many
982 millions of years (Koonin et al. 2008).
983
984 **Non-phylogenetic Interaction**

985 A non-phylogenetic interaction was detected. This interaction represents variation in
986 prevalence caused by specific host-virus pairings having prevalences beyond that
987 which would be expected by the simple addition of the individual host and virus
988 means. A non-phylogenetic interaction could be caused by a large range of factors,
989 some biological and some due to the specifics of the model, many of which would be
990 likely to be acting simultaneously to generate this signal. One possibility is
991 coevolution between the host and virus that occurred after both diverged from their
992 common ancestor with the closest related species in the study. Another is the
993 complete absence of coevolution, where spillover from a primary or group of primary
994 hosts causes either a constant very low prevalence of dead-end infections, which are
995 none-the-less detectable by PCR. Related to this is a statistical issue involving cases
996 where not every species in the study is within a virus' host range, and the species that
997 are within the host range are not closely related. In this case, the variation is not
998 absorbed by the host evolutionary interaction and almost no host has a prevalence
999 close to the mean across hosts, as in many species the prevalence is zero, which
1000 causes the mean to be considerably lower than the average prevalence in the species
1001 the virus does infect. This effect would be magnified if the sampling occurs during an
1002 epizootic. More broadly, anything that changes the epidemiological parameters of a
1003 virus in a specific host will lead to a non-phylogenetic interaction. Considering the
1004 variation in the natural history of viruses and the lesser, but still significant, variation
1005 in the natural history of bumblebees, a large non-phylogenetic interaction is to be
1006 expected.
1007
1008 **Conclusion**

1009 While it is clear that viruses are abundant in pollinators, the factors that determine the
1010 distribution of pollinator viruses have remained uncertain, outside of a few well-
1011 studied cases (Fürst et al. 2014; McMahon et al. 2015). With the novel viruses
1012 discovered in this study, we have investigated predictors of these virus/host
1013 associations and found that both the host evolutionary history and the identity of the
1014 virus contributes to this distribution. This supports both theory and prior empirical
1015 evidence that related species are more at risk of infection from each other's diseases
1016 than the diseases of distantly related species. However, the importance of the viral
1017 identity and unique interactions between host-virus pairs suggests that the
1018 introduction of a novel virus into a community is likely to have unpredictable effects
1019 even when no close relatives of currently known hosts are present. This highlights the
1020 risk posed by disease spillover for the conservation not only of wild pollinator
1021 communities, but also to communities consisting of related animal or plant species in
1022 general.
1023

1024 **Data Availability**

1025 The data and code for running the analyses is available on github under a GPLv3
1026 licence, as code uses code taken from other GPLv3 licenced works
1027 (https://github.com/dpascall/bumblebee-virus-cophylo).
1028

1029 **Acknowledgements**

1030 We thank Jarrod Hadfield for extensive statistical advice, modifications to
1031 MCMCglmm and for helpful comments on this manuscript, Dave Goulson for
1032 assistance identifying some specimens and to Ben Longdon and Bethany Clark for
1033 further comments. Rowan Doff assisted in the lab with RFLP analyses and Claire

1034 Webster with DNAse treatments. Claire Webster, Jarrod Hadfield and Florian Bayer
1035 helped with fieldwork.
1036
1037 **Funding**
1038 This work was funded by a BBSRC SWBIO DTP PhD stipend to DP, a Royal Society
1039 Dorothy Hodgkin Fellowship to LW, Wellcome Trust Research Career Development
1040 Fellowship WT085064 to DJO.
1041
1042 **References**
1043 Altschul, S.F. et al., 1990. Basic local alignment search tool. *Journal of Molecular*
1044 *Biology*, 215(3), pp.403–410.
1045 Arbulo, N, Santos, E, Salvarrey, S, & Invernizzi, C., 2011. Proboscis length and
1046 resource utilization in two ruguayan bumblebees: Bombus atratus Franklin and
1047 Bombus bellicosus Smith (Hymenoptera: Apidae). *Neotropical Entomology*,
1048 40(4), pp.483–488.
1049 Arneberg, P. et al., 1998. Host densities as determinants of abundance in parasite
1050 communities. *Proceedings of the Royal Society B: Biological Sciences*,
1051 265(1403), pp.1283–1289.
1052 Bailey, L. & Ball, B. V., 1991. *Honey Bee Pathology* 2nd Editio., London: Academic
1053 Press Inc.
1054 Bailey, L., Gibbs, A.J. & Woods, R.D., 1963. Two viruses from adult honey bees
1055 (Apis mellifera Linnaeus). *Virology*, 21(3), pp.390–395.
1056 Bailey, L. & Woods, R.D., 1974. Three previously undescribed viruses from the
1057 honey bee. *Journal of General Virology*, 25(2), pp.175–186.
1058 Bailey, L. & Woods, R.D., 1977. Two more small RNA viruses from honey bees and

1059 further observations on sacbrood and acute bee-paralysis viruses. *Journal of*
1060 *General Virology*, 37(1), pp.175–182.
1061 Baker, C. & Antonovics, J., 2012. Evolutionary determinants of genetic variation in
1062 susceptibility to infectious diseases in humans. *PLoS ONE*.
1063 Bandín, I. & Dopazo, C.P., 2011. Host range, host specificity and hypothesized host
1064 shift events among viruses of lower vertebrates. *Veterinary Research*, 42(67).
1065 Boccardo, G. et al., 1985. Three seedborne cryptic viruses containing double-stranded
1066 RNA isolated from white clover. *Virology*, 147(1), pp.29–40.
1067 Bouckaert, R. et al., 2014. BEAST 2: A Software Platform for Bayesian Evolutionary
1068 Analysis A. Prlic, ed. *PLoS Computational Biology*, 10(4), p.e1003537.
1069 Bouckaert, R. & Drummond, A., 2015. bModelTest: Bayesian phylogenetic site
1070 model averaging and model comparison. *bioRxiv*, p.020792.
1071 Brennecke, J. et al., 2007. Discrete Small RNA-Generating Loci as Master Regulators
1072 of Transposon Activity in Drosophila. *Cell*, 128(6), pp.1089–1103.
1073 Bronkhorst, A.W. et al., 2012. The DNA virus Invertebrate iridescent virus 6 is a
1074 target of the Drosophila RNAi machinery. *Proceedings of the National Academy*
1075 *of Sciences*, 109(51), pp.E3604–E3613.
1076 Buchfink, B., Xie, C. & Huson, D.H., 2015. Fast and sensitive protein alignment
1077 using DIAMOND. *Nature methods*, 12(1), pp.59–60.
1078 Cameron, S.A., Hines, H.M. & Williams, P.H., 2007. A comprehensive phylogeny of
1079 the bumble bees (Bombus). *Biological Journal of the Linnean Society*, 91(1),
1080 pp.161–188.
1081 Carpenter, B. et al., 2017. Stan: A probabilistic programming language. *Journal of*
1082 *Statistical Software*, 76(1), pp.1–32.
1083 Chandra, R.K., 1983. Nutrition, immunity, and infection: Present knowledge and

1084 future directions. *The Lancet*, 321(8326), pp.688–691.
1085 Chiba, S. et al., 2011. Widespread endogenization of genome sequences of non-
1086 retroviral RNA viruses into plant genomes. *PLoS Pathogens*, 7(7).
1087 Civitello, D.J. et al., 2015. Biodiversity inhibits parasites: Broad evidence for the
1088 dilution effect. *Proceedings of the National Academy of Sciences*, 112(28),
1089 pp.8667–8671.
1090 Cooper, N. et al., 2012. Phylogenetic host specificity and understanding parasite
1091 sharing in primates. *Ecology Letters*, 15(12), pp.1370–1377.
1092 Cornman, R.S. et al., 2012. Pathogen webs in collapsing honey bee colonies. *PLoS*
1093 *ONE*, 7(8), pp.1–21.
1094 Curtis, V.A., 2014. Infection-avoidance behaviour in humans and other animals.
1095 *Trends in Immunology*, 35(10), pp.457–464.
1096 Davies, T.J. & Pedersen, A.B., 2008. Phylogeny and geography predict pathogen
1097 community similarity in wild primates and humans. *Proceedings of the Royal*
1098 *Society B: Biological Sciences*, 275(1643), pp.1695–1701.
1099 Deleris, A. et al., 2006. Hierarchical action and inhibition of plant dicer-like proteins
1100 in antiviral defense. *Science*, 313(5783), pp.68–71.
1101 Drummond, A.J. et al., 2012. Bayesian phylogenetics with BEAUti and the BEAST
1102 1.7. *Molecular Biology and Evolution*, 29(8), pp.1969–1973.
1103 Drummond, A.J. et al., 2006. Relaxed phylogenetics and dating with confidence D.
1104 Penny, ed. *PLoS Biology*, 4(5), pp.699–710.
1105 Durrer, S. & Schmid-Hempel, P., 1994. Shared Use of Flowers Leads to Horizontal
1106 Pathogen Transmission. *Proceedings of the Royal Society B: Biological*
1107 *Sciences*, 258(1353), pp.299–302.
1108 Ebert, T.A., Brlansky, R. & Rogers, M., 2010. Reexamining the Pooled Sampling

1109 Approach for Estimating Prevalence of Infected Insect Vectors. *Annals of the*
1110 *Entomological Society of America*, 103(6), pp.827–837.
1111 Fauquet, C.M. & Stanley, J., 2005. Revising the way we conceive and name viruses
1112 below the species level: A review of geminivirus taxonomy calls for new
1113 standardized isolate descriptors. *Arch Virol*, 150, pp.2151–2179.
1114 Félix, M.A. et al., 2011. Natural and experimental infection of Caenorhabditis
1115 nematodes by novel viruses related to nodaviruses J. Hodgkin, ed. *PLoS Biology*,
1116 9(1), p.e1000586.
1117 Fenton, A. et al., 2015. Are All Hosts Created Equal? Partitioning Host Species
1118 Contributions to Parasite Persistence in Multihost Communities. *The American*
1119 *Naturalist*, 186(5), pp.610–622.
1120 Ferreira, M.A.R. & Suchard, M.A., 2008. Bayesian analysis of elapsed times in
1121 continuous-time Markov chains. *Canadian Journal of Statistics*, 36(3), pp.355–
1122 368.
1123 Finn, R.D. et al., 2014. Pfam: The protein families database. *Nucleic Acids Research*,
1124 42(D1), pp.D222–D230.
1125 Fürst, M. a. et al., 2014. Disease associations between honeybees and bumblebees as a
1126 threat to wild pollinators. *Nature*, 506(7488), pp.364–366.
1127 Garibaldi, L.A. et al., 2013. Wild pollinators enhance fruit set of crops regardless of
1128 honey bee abundance. *Science*, 340(6127), pp.1608–1611.
1129 Gibbs, A.J. & Gower, J.C., 1960. The use of a multiple-transfer method in plant virus
1130 transmission studies - Some statistical points arising in the analysis of results.
1131 *Annals of Applied Biology*, 48(1), pp.75–83.
1132 Goodwin, S., 1995. Seasonal phenology and abundance of early-, mid-and long-
1133 season bumble bees in southern England. *Journal of Apicultural Research*, 34(2),

1134 pp.79–87.
1135 Gorbalenya, A.E. et al., 2002. The palm subdomain-based active site is internally
1136 permuted in viral RNA-dependent RNA polymerases of an ancient lineage.
1137 *Journal of Molecular Biology*, 324(1), pp.47–62.
1138 Goulson, D. et al., 2005. Causes of rarity in bumblebees. *Biological Conservation*,
1139 122(1), pp.1–8.
1140 Goulson, D. & Darvill, B., 2004. Niche overlap and diet breadth in bumblebees; are
1141 rare species more specialized in their choice of flowers? *Apidologie*, 35(1),
1142 pp.55–63.
1143 Goulson, D., Lye, G.C. & Darvill, B., 2008. Diet breadth, coexistence and rarity in
1144 bumblebees. *Biodiversity and Conservation*, 17(13), pp.3269–3288.
1145 Grabherr, M.G. et al., 2011. Full-length transcriptome assembly from RNA-Seq data
1146 without a reference genome. *Nature Biotechnology*, 29(7), pp.644–652.
1147 Graystock, P., Goulson, D. & Hughes, W.O.H., 2015. Parasites in bloom: flowers aid
1148 dispersal and transmission of pollinator parasites within and between bee
1149 species. *Proceedings of the Royal Society B-Biological Sciences*, 282(1813),
1150 p.20151371.
1151 Hadfield, J.D. et al., 2014. A tale of two phylogenies: comparative analyses of
1152 ecological interactions. *The American Naturalist*, 183(2), pp.174–87.
1153 Harder, L.D., 1985. Morphology as a predictor of flower choice by bumble bees.
1154 *Ecology*, 66(1), pp.198–210.
1155 Harmon-Threatt, A.N. & Ackerly, D.D., 2013. Filtering across Spatial Scales:
1156 Phylogeny, Biogeography and Community Structure in Bumble Bees. *PLoS*
1157 *ONE*, 8(3), p.e60446.
1158 Heesterbeek, J.A.P., 2002. A brief history of R0 and a recipe for its calculation. *Acta*

1159 *Biotheoretica*, 50, pp.189–204.
1160 Heled, J. & Drummond, A.J., 2012. Calibrated tree priors for relaxed phylogenetics
1161 and divergence time estimation. *Systematic Biology*, 61(1), pp.138–149.
1162 Henikoff, S. & Henikoff, J.G., 1992. Amino acid substitution matrices from protein
1163 blocks. *Proceedings of the National Academy of Sciences of the United States of*
1164 *America*, 89(22), pp.10915–10919.
1165 Hines, H.M., 2008. Historical Biogeography, Divergence Times, and Diversification
1166 Patterns of Bumble Bees (Hymenoptera: Apidae: Bombus). *Systematic Biology*,
1167 57(1), pp.58–75.
1168 Holmes, E.C., 2003. Molecular Clocks and the Puzzle of RNA Virus Origins. *Journal*
1169 *of Virology*, 77(7), pp.3893–3897.
1170 Huang, S. et al., 2014. Phylogenetically related and ecologically similar carnivores
1171 harbour similar parasite assemblages. *Journal of Animal Ecology*, 83(3), pp.671–
1172 680.
1173 Hueffer, K. et al., 2003. The Natural Host Range Shift and Subsequent Evolution of
1174 Canine Parvovirus Resulted from Virus-Specific Binding to the Canine
1175 Transferrin Receptor The Natural Host Range Shift and Subsequent Evolution of
1176 Canine Parvovirus Resulted from Virus-Specific Bind. *Journal of Virology*,
1177 77(3), pp.1718–1726.
1178 Inouye, D.W., 1978. Resource Partitioning in Bumblebees: Experimental Studies of
1179 Foraging Behavior. *Ecology*, 59(4), pp.672–678.
1180 Johnson, M.B. et al., 2011. Parasite transmission in social interacting hosts:
1181 Monogenean epidemics in guppies. *PLoS ONE*, 6(8).
1182 Kamer, G. & Argos, P., 1984. Primary structural comparison of RNA-dependent
1183 polymerases from plant, animal and bacterial viruses. *Nucleic Acids Research*,

1184 12(18), pp.7269–7282.
1185 Kanehisa, M. et al., 2002. The KEGG databases at GenomeNet. *Nucleic Acids*
1186 *Research*, 30(1), pp.42–6.
1187 Katoh, K. et al., 2005. MAFFT version 5: improvement in accuracy of multiple
1188 sequence alignment. *Nucleic Acids Research*, 33(2), pp.511–518.
1189 Katoh, K., Rozewicki, J. & Yamada, K.D., 2017. MAFFT online service: multiple
1190 sequence alignment, interactive sequence choice and visualization. *Briefings in*
1191 *Bioinformatics*.
1192 Katzourakis, A. & Gifford, R.J., 2010. Endogenous viral elements in animal genomes.
1193 *PLoS Genetics*, 6(11), p.e1001191.
1194 Koloniuk, I., Přibylová, J. & Fránová, J., 2018. Molecular characterization and
1195 complete genome of a novel nepovirus from red clover. *Archives of Virology*,
1196 163(5), pp.1387–1389.
1197 Koonin, E. V., Dolja, V. V. & Krupovic, M., 2015. Origins and evolution of viruses
1198 of eukaryotes: The ultimate modularity. *Virology*, 479–480, pp.2–25.
1199 Koonin, E. V et al., 2008. The Big Bang of picorna-like virus evolution antedates the
1200 radiation of eukaryotic supergroups. *Nature Reviews Microbiology*, 6(12),
1201 pp.925–939.
1202 Langmead, B. & Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2.
1203 *Nature Methods*, 9(4), pp.357–359.
1204 Langsjoen, R.M. et al., 2018. Chikungunya virus strains show lineage-specific
1205 variations in virulence and cross-protective ability in murine and nonhuman
1206 primate models. *mBio*.
1207 Lewis, S.H. et al., 2018. Pan-arthropod analysis reveals somatic piRNAs as an
1208 ancestral defence against transposable elements. *Nature Ecology and Evolution*,

1209 2(1), pp.174–181.
1210 Liu, Y., Gelman, A. & Zheng, T., 2015. Simulation-efficient shortest probability
1211 intervals. *Statistics and Computing*, 25(4), pp.809–819.
1212 Longdon, B. et al., 2011. Host phylogeny determines viral persistence and replication
1213 in novel hosts. *PLoS Pathogens*, 7(9), p.e1002260.
1214 Longdon, B. et al., 2018. Host shifts result in parallel genetic changes when viruses
1215 evolve in closely related species. *PLoS Pathogens*.
1216 Longdon, B. et al., 2015. The causes and consequences of changes in virulence
1217 following pathogen host shifts. *PLoS Pathogens*, 11(3), p.e1004728.
1218 Longdon, B. et al., 2014. The Evolution and Genetics of Virus Host Shifts. *PLoS*
1219 *Pathogens*, 10(11), p.e1004395.
1220 Lye, G.C. et al., 2010. Forage use and niche partitioning by non-native bumblebees in
1221 New Zealand: Implications for the conservation of their populations of origin.
1222 *Journal of Insect Conservation*, 14(6), pp.607–615.
1223 Manley, R., Boots, M. & Wilfert, L., 2015. Emerging viral disease risk to pollinating
1224 insects: Ecological, evolutionary and anthropogenic factors. *Journal of Applied*
1225 *Ecology*, 52(2), pp.331–340.
1226 McArt, S.H. et al., 2014. Arranging the bouquet of disease: Floral traits and the
1227 transmission of plant and animal pathogens. *Ecology Letters*, 17(5), pp.624–636.
1228 McMahon, D.P. et al., 2015. A sting in the spit: Widespread cross-infection of
1229 multiple RNA viruses across wild and managed bees. *Journal of Animal*
1230 *Ecology*, 84(3), pp.615–624.
1231 van Mierlo, J.T. et al., 2012. Convergent Evolution of Argonaute-2 Slicer Antagonism
1232 in Two Distinct Insect RNA Viruses. *PLoS Pathogens*, 8(8), p.e1002872.
1233 de Miranda, J.R. et al., 2015. Genome characterization, prevalence and distribution of

1234 a macula-like virus from Apis mellifera and Varroa destructor. *Viruses*, 7(7),
1235 pp.3586–3602.
1236 Mordecai, G.J. et al., 2015. Diversity in a honey bee pathogen: first report of a third
1237 master variant of the Deformed Wing Virus quasispecies. *The ISME Journal*,
1238 10(5), pp.1–10.
1239 Murray, T.E. et al., 2008. Cryptic species diversity in a widespread bumble bee
1240 complex revealed using mitochondrial DNA RFLPs. *Conservation Genetics*,
1241 9(3), pp.653–666.
1242 Pei, J., Kim, B.H. & Grishin, N. V., 2008. PROMALS3D: A tool for multiple protein
1243 sequence and structure alignments. *Nucleic Acids Research*, 36(7), pp.2295–
1244 2300.
1245 Rambaut, A. et al., 2017. Tracer v1.7. Available at: http://beast.bio.ed.ac.uk/Tracer.
1246 Rao, S., Poinar, G. & Henley, D., 2017. A scientific note on rare parasitism of the
1247 bumble bee pollinator, Bombus impatiens, by a mermithid nematode,
1248 Pheromermis sp. (Nematoda: Mermithidae). *Apidologie*, 48(1), pp.75–77.
1249 Remnant, E.J. et al., 2017. A Diverse Range of Novel RNA Viruses in Geographically
1250 Distinct Honey Bee Populations. *Journal of Virology*, (May), p.JVI.00158-17.
1251 van Rij, R.P. et al., 2006. The RNA silencing endonuclease Argonaute 2 mediates
1252 specific antiviral immunity in Drosophila melanogaster. *Genes and*
1253 *Development*, 20(21), pp.2985–2995.
1254 van Riper, C. et al., 1986. The Epizootiology and ecological significance of malaria in
1255 Hawaiian land birds. *Ecological Monographs*, 56(4), pp.327–344.
1256 Roberts, J.M.K., Anderson, D.L. & Durr, P.A., 2018. Metagenomic analysis of
1257 Varroa-free Australian honey bees (Apis mellifera) shows a diverse
1258 Picornavirales virome. *Journal of General Virology*, 99, pp.818–826.

1259 Roberts, K. et al., 2018. Changes in temperature alter susceptibility to a virus
1260 following a host shift. *bioRxiv*.
1261 Ruiz-González, M.X. et al., 2012. Dynamic transmission, host quality, and population
1262 structure in a multihost parasite of bumblebees. *Evolution*, 66(10), pp.3053–
1263 3066.
1264 Runckel, C. et al., 2011. Temporal analysis of the honey bee microbiome reveals four
1265 novel viruses and seasonal prevalence of known viruses, Nosema, and Crithidia.
1266 *PLoS ONE*, 6(6), p.e20656.
1267 Salathé, R.M. & Schmid-Hempel, P., 2011. The Genotypic Structure of a Multi-Host
1268 Bumblebee Parasite Suggests a Role for Ecological Niche Overlap. *PLOS ONE*,
1269 6(8), p.e22054.
1270 Schoonvaere, K. et al., 2018. Study of the Metatranscriptome of Eight Social and
1271 Solitary Wild Bee Species Reveals Novel Viruses and Bee Parasites. *Frontiers in*
1272 *Microbiology*, 9, p.177.
1273 Schoonvaere, K. et al., 2016. Unbiased RNA Shotgun Metagenomics in Social and
1274 Solitary Wild Bees Detects Associations with Eukaryote Parasites and New
1275 Viruses R. Lu, ed. *PLOS ONE*, 11(12), p.e0168456.
1276 Shi, M. et al., 2016. Redefining the invertebrate RNA virosphere. *Nature*, 540, pp.1–
1277 12.
1278 Smith, K.M. & Markham, R., 1944. Two new viruses affecting tobacco and other
1279 plants. *Phytopathology*, 34, pp.324–329.
1280 Stan Development Team, 2017. *Stan Modeling Language: User's Guide and*
1281 *Reference Manual* 2.17.0.
1282 Streicker, D.G. et al., 2012. Variable evolutionary routes to host establishment across
1283 repeated rabies virus host shifts among bats. *Proceedings of the National*

1284 *Academy of Sciences*.
1285 Suttle, C.A., 2007. Marine viruses--major players in the global ecosystem. *Nature*
1286 *Reviews Microbiology*, 5(10), pp.801–812.
1287 Suzuki, Y. et al., 2017. Uncovering the repertoire of endogenous flaviviral elements
1288 in Aedes. *Journal of Virology*.
1289 Thompson, K.H., 1962. Estimation of the Proportion of Vectors in a Natural
1290 Population of Insects. *Biometrics*, 18(4), pp.568–578.
1291 Untergasser, A. et al., 2012. Primer3-new capabilities and interfaces. *Nucleic Acids*
1292 *Research*, 40(15), p.e115.
1293 Vanbergen, A.J. & the Insect Pollinators Initiative, 2013. Threats to an ecosystem
1294 service: Pressures on pollinators. *Frontiers in Ecology and the Environment*,
1295 11(5), pp.251–259.
1296 Wang, X.-B. et al., 2011. The 21-Nucleotide, but Not 22-Nucleotide, Viral Secondary
1297 Small Interfering RNAs Direct Potent Antiviral Defense by Two Cooperative
1298 Argonautes in Arabidopsis thaliana. *The Plant Cell*, 23(4), pp.1625–1638.
1299 Waxman, D. et al., 2014. Inferring Host Range Dynamics from Comparative Data:
1300 The Protozoan Parasites of New World Monkeys. *The American Naturalist*,
1301 184(1), pp.65–74.
1302 Webster, C.L. et al., 2015. The discovery, distribution, and evolution of viruses
1303 associated with drosophila melanogaster. *PLoS Biology*, 13(7), p.e1002210.
1304 Williams, P.H.P.H.P.H. & Osborne, J.L., 2009. Bumblebee vulnerability and
1305 conservation world-wide. *Apidologie*, 40(3), pp.367–387.
1306 Woolhouse, M.E.J. & Gowtage-Sequeria, S., 2005. Host range and emerging and
1307 reemerging pathogens. *Emerging Infectious Diseases*, 11(12), pp.1842–1847.
1308 Zanotto, P.M. et al., 1996. A reevaluation of the higher taxonomy of viruses based on

1309 RNA polymerases. *Journal of Virology*, 70(9), pp.6083–6096.

1310