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

Short title: Evolutionary signals in bumblebee-virus interaction networks

David J. Pascall[1,#a*], Matthew C. Tinsley[2], Darren J. Obbard[3,4] and Lena Wilfert[1,#b]

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

[2] Biological and Environmental Sciences, Stirling University, Stirling, FK9 4LA, UK

[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

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

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

[*] Corresponding author

Email: david.pascall@glasgow.ac.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. We first searched for viruses using a metatranscriptomic approach, finding at least 30 new viruses of bumblebees, and assayed a subset of them for their presence and absence in different host 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 having different prevalences for a subset of these viruses. While much of the variation in the prevalence of the viruses can be explained by the idiosyncrasies of individual bumblebee-virus pairings, there is a phylogenetic signal with related bumblebee species being infected at similar frequencies by the same sets of viruses. 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 of infection.

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 [1]. One especially important group of pathogens are the viruses, whose ubiquity leads them to have a disproportionate role in the regulation of natural populations [2]. Viruses are relevant in populations that humans manage for

economic and conservation reasons, such as bumblebees, which are both in decline [3] and important providers of ecosystem services [4].

Bumblebees, genus *Bombus*, are a primitively eusocial group of important wild pollinators. Many bumblebee species have experienced population declines, linked to biotic and abiotic stressors such as habitat degradation, pesticide use and shared infectious diseases for example caused by viral pathogens [5]. In contrast to honeybee viruses, which have been intensively studied, and have in many cases been found to represent multihost pathogens (see Manley et al (2015) [6] and the references within), bumblebee-specific viruses are comparatively poorly studied, and it is unknown how widely they are shared between host species.

For a species to be a multihost pathogen, some degree of opportunity for cross-species transmission must exist. Our definition of multihost pathogens follows that of Fenton et al. [7]. As such, multihost pathogens are defined to include two conceptually distinct groups: 'facultative multihost pathogens' that are able to maintain transmission chains in multiple host species (i.e. R_0>=1 in multiple host species) and 'obligate multihost pathogens', which rely on sufficiently high rates of cross-species transmission to offset unsustainable transmission within individual host species (i.e. 0<R_0<1 within host species, R_0>=1 overall). In addition, pathogens that maintain transmission in a single host (R_0>=1) but experience regular spillover (with or without the expectation of onward transmission: 0<= R_0<1) are included as being effectively multihost pathogens within our definition. R_0 is defined as the expected number of secondary infections caused by a single typical infected individual in an entirely naïve host population [8]. We define cross-species transmission as the movement of a multihost pathogen between host species within its host range. This contrasts with

99 host shifting, which we define as a transmission event to a new host species, leading
100 to a change in host range. However there is necessarily some unavoidable ambiguity
101 between cross-species transmission and host shifting in the case of pathogens that
102 exhibit rare spillover events.
103
104 The opportunity for cross-species transmission, which explains the large number of
105 viruses originally detected in honeybees that are present in bumblebees, may be
106 created by niche overlap in foraging [9]. Bumblebee nests are provisioned by foraging
107 workers who gather pollen and nectar from flowers in the surrounding area.
108 Considerable interspecific differences in plant species utilization by foragers of
109 different species are commonly observed [10-13], but this is not a universal
110 phenomenon [14], and the degree of overlap may depend on the diversity of flowers
111 currently in bloom. In bumblebees, flower choice of foragers is correlated with
112 species tongue length [12,13], which implicitly incorporates shared behavioural
113 characteristics between closely related bumblebee species as there is phylogenetic
114 correlation between tongue length and relatedness [15]. Different species of
115 bumblebee also exhibit incomplete temporal separation throughout the year, causing
116 some degree of partitioning in niche space even when they are spatially sympatric
117 [16]. This ecology leads to a complex interaction network between bumblebee
118 species, as well as sympatric honeybees, which may structure cross-species
119 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 [17]. Incompatibility can break down if evolution in the pathogen
137 or host removes the physiological or molecular barriers to infection, as shown when
138 Canine parvovirus type 2 emerged from Feline panleukopenia virus after gaining the
139 ability to bind to canine transferrin receptors [18].
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 [19], population densities [20], the biodiversity of the community [21]
145 and host avoidance behaviours [22]. Variation in infection risk among host species
146 can also be driven by physiological and molecular factors, with hosts having varying
147 suitability for the replication of a given parasite. In the extreme case, a host species
148 may exhibit condition-dependent susceptibility; where infection can only occur when

149 the immune system is suppressed, either directly, through an immunosuppressant
150 disease or chemical agent, or indirectly, through trade-offs in resource allocation
151 brought about by malnutrition [23]. Both behavioural and ecological factors, leading
152 to differences in contact rate, and physiological factors, leading to differences in
153 infection probability on contact, may be phylogenetically correlated [15,24].
154
155 **-------**
156 **Box 1 – Definition of Terms**
157 Co-phylogenetic generalized linear mixed models that incorporate phylogenetic
158 variance from multiple clades [25,26] have been used relatively rarely, and a
159 biological interpretation of the model terms may not be immediately familiar. In the
160 host-parasite context, this approach can be used to model how the probability of
161 infection is predicted by both host and parasite species, allowing for covariance
162 induced by the relationships within each group, and the interactions between these
163 model terms. This can be considered either at the species-wide level (i.e. the
164 probability that infection will occur at all in a given host/parasite pairing), or at the
165 level of individuals within species (i.e. infection prevalence). Here we provide verbal
166 descriptions of how the terms can be interpreted, as well as references to a figure in
167 Hadfield et al. (2014) [25] where each of these effects is illustrated graphically:
168
169 *Phylogenetic Effect*: Variation in the mean value of a trait among species that is
170 explained by phylogenetic divergence. For example, more closely-related hosts might
171 be more similar in susceptibility to viral infection (display higher viral prevalence),
172 irrespective of virus species. Equivalently, more closely-related viruses might be
173 more similar in infectivity, irrespective of host species (Figs 1a and b in Hadfield et

174 al. (2014) [25]).

175

176 *Species Effect*: Variation in the mean value of a trait among species that is
177 not explained by a *Phylogenetic Effect*. For example, much of the variation in
178 prevalence among viral species (irrespective of host), may not be explained by the
179 virus phylogeny but instead depend on lineage-specific viral traits (Figs 1g and f in
180 Hadfield et al. (2014) [25]).

181

182 *Non-phylogenetic Interaction*: The interaction term between host and parasite *Species*
183 *Effects*, such that variation in the mean value of a trait within particular host/parasite
184 pairings depends the specifics of the host and parasite involved in a way not affected
185 by their evolutionary divergence. For example, variation in prevalence between
186 particular pairings that is caused by the interaction between lineage-specific host and
187 viral traits (Fig 1h in Hadfield et al. (2014) [25]).

188

189 *Coevolutionary Interaction*: The interaction term between host and parasite
190 *Phylogenetic Effects*, such that variation in the mean value of a trait within particular
191 host/parasite pairings depends on the evolutionary divergence among species in both
192 host and parasite clades. For example, if the prevalence of infection is more similar
193 among pairings of closely-related hosts and closely-related parasites than would be
194 expected from the host and parasite phylogenies and species-means alone. (Fig 1e in
195 Hadfield et al. (2014) [25]).

196

197 *Evolutionary Interaction*: The interaction term between the host (or parasite)
198 *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. (Fig 1c and d in Hadfield et al. (2014) [25]).

A new multihost virus can arise in two ways, either through a virus species gaining a new host [27] (Fig 1a), or through a speciation event in a multihost virus (i.e. giving rise to two sibling multihost virus species; Fig 1b). When novel multihost viruses are generated through host shifting (Fig 1a), a 'host evolutionary interaction' effect (see Box 1) 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 (Fig 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.

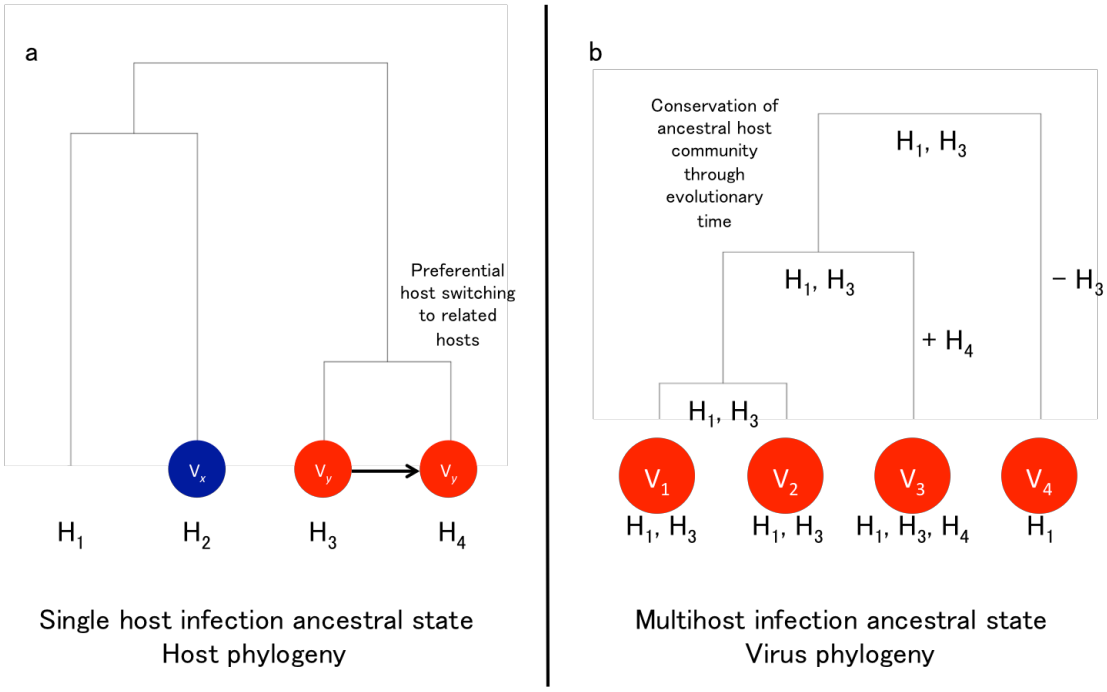


Fig 1 Mechanisms for the generation of novel multihost viruses. The generation of novel multihost viruses through host shifting (1a) leads to a 'host evolutionary interaction' effect (Box1), as the consistent switching of viruses (V) to hosts (H) closely related to their ancestral host will lead to related 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.

Here we test 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 [28], Acute bee paralysis virus [29] and Hubei partiti-like virus 34 [30,31]. We analysed virus prevalence using co-phylogenetic models to determine the presence or absence and relative strengths of the phylogenetic signals that are expected to shape the host/virus assemblage in this system, and performed tests to attempt to determine the mechanisms driving this.

240

241 **Methods**

242 **Sampling strategy**

243 A total of 926 individual bumblebees of 13 species were collected on the wing from
244 nine sites across Scotland in July and August of 2009 and 2011, and frozen in liquid
245 nitrogen or at -80°C. In 2009, we sampled the Ochil Hills, Glenmore, Dalwhinnie,
246 Stirling, Iona, Staffa, and the Pentlands, and in 2011 we sampled Edinburgh and
247 Gorebridge (Supp Table 1). The cryptic species complex of *Bombus terrestris*,
248 *Bombus lucorum, Bombus cryptarum* and *Bombus magnus* was resolved using RFLP
249 analysis following Murray et al. (2008) [32]. All individuals were bisected
250 longitudinally prior to RNA extraction. One half of each bumblebee was used in
251 pooled RNA extractions of 2-11 individuals per species (median 10; Supp Table 2).
252 Two of these pools ('DIV' and 'P11') were included in the RNAseq, but excluded
253 from prevalence testing. The groups of bumblebees were ground in liquid nitrogen
254 and added to TRIzol reagent (Life Technologies) for RNA extractions, following the
255 manufacturer's standard protocol. The RNA concentrations in the pooled samples
256 were equalized to approximately 200 ng/ul/individual based on Nanodrop
257 measurements.

258

259 **RNA Sequencing and Bioinformatics**

260 The RNA was combined by species for *B. terrestris* (239 individuals), *Bombus*
261 *pascuorum* (212 individuals), *B. lucorum* (182 individuals) and other *Bombus* (293
262 individuals) into four large RNA pools. These large pools were sequenced using the
263 Illumina HiSeq platform with 100bp paired end reads (Beijing Genomics Institute)
264 after poly-A selection. This excludes ribosomal and bacterial RNA, and will enrich

265 for mRNAs and those RNA viruses that have polyadenylated genomes or products.
266 The single-species bumblebee pools were subsequently re-sequenced following
267 duplex specific nuclease normalization, to reduce rRNA representation, and enrich for
268 rare transcripts while retaining non-polyadenylated viruses and products. The small
269 RNAs of the same RNA pools of *B. terrestris*, *B. lucorum* and *B. pascuorum* were
270 also sequenced to test for the replication of viruses identified via the transcriptome
271 sequencing.
272
273 For each pool, paired end RNAseq data were initially mapped to the published
274 *Bombus terrestris* and *B. impatiens* genomes using bowtie2 [33] to reduce the
275 representation of conserved bumblebee sequences. Read pairs that did not map
276 concordantly, including divergent bumblebee sequences and other associated
277 microbiota, were assembled *de novo* using Trinity 2.2.0 [34] as paired end libraries,
278 following automated trimming ('--trimmomatic') and digital read normalisation ('--
279 normalize_reads'). Where two RNAseq libraries (Poly-A and DSN) had been
280 sequenced, these were combined for assembly.
281
282 To identify putative viruses, all long open reading frames from each contig were
283 identified and concatenated to provide a 'bait' sequence for similarity searches using
284 Diamond [35] and BLASTp [36]. Contigs shorter than 500 base pairs were discarded.
285 These contig translations were used to search against a Diamond database comprising
286 all of the virus protein sequences available in NCBI database 'nr', and all of the
287 Dipteran, Hymenopteran, Nematode, Fungal, Protist, and prokaryotic proteins
288 available in NCBI database 'refseq_protein' (mode 'blastp'; e-value 0.001; maximum
289 of one match). Matches to phage and short matches to large DNA viruses were

290 excluded. Remaining contigs were manually curated to identify and annotate high-
291 confidence virus-like sequences. To quantify approximate fold-coverage, and to
292 assess viRNA properties, the raw RNAseq and trimmed small RNA reads were
293 mapped against the putative viral contigs using bowtie2's '--very-sensitive' setting
294 and retaining only the top map [33], from this we recorded the number of mapped
295 reads per kilobase of transcript per million mapped reads. We considered viruses to
296 show strong evidence of replication in the host if they had at least 50 mapping siRNA
297 reads with a size distribution sharply peaked at 22nt (viRNAs are generated from
298 replicating viruses by Dcr2). Following Fauquet and Stanley (2005) [37], we
299 designated contigs exhibiting less than 90% nucleotide identity as separate viruses
300 and those exhibiting greater than or equal to 90% identity as strains of known viruses.
301

302 **PCR Validation and Testing**

303 A subset of contigs were chosen for manual validation. All chosen contigs met both of
304 the following conditions: the presence of mapping reads in the bumblebee small
305 RNAs (for the *B. terrestris*, *B. pascuorum* and *B. lucorum* pools; not a condition for
306 the mixed *Bombus* pool) or the transcriptomic RNAs (for the other *Bombus* pool
307 where small RNAs were not generated), and the closest blast match being viral RNA-
308 dependent RNA-polymerase. Internal primers for these contigs were generated using
309 primer3 [38] and amplification of the target was verified via Sanger sequencing. See
310 Supp Table 3 for PCR conditions and primer sequences. Mayfield virus 1 and 2 were
311 Sanger sequence validated over the entirety of the contig. The Loch Morlich and
312 River Liunaeg virus sequences were generated by the connection of several disjoint
313 contigs by Sanger sequencing. Black Hill virus was excluded from further analysis as

314 it was found that that the PCR reaction amplified a host sequence that could not be
315 visually differentiated from the virus product.
316
317 **Phylogenetic Inference**
318 Following Cameron et al. (2007) [39], we inferred the bumblebee phylogeny using
319 cytochrome oxidase I, elongation factor 1-alpha, opsin, phosphoenolpyruvate
320 carboxykinase, 16S and arginine kinase genes. To break up long branches and allow
321 dating, additional species not sampled in the field were added from public databases
322 (see Supp Table 4 for genbank accession numbers and species included). The DNA
323 sequences were aligned with MAFFT using the L-INS-i setting [40,41]. The 6 gene
324 alignments were then used to generate the phylogeny in BEAST v2.4.5 [42], treating
325 each file as a separate partition, using bModelTest [43] with the
326 'transitionTransversion split' setting and a calibrated Yule tree prior [44]. An
327 uncorrelated lognormal relaxed clock was fitted to each partition, with exponential
328 (λ=1) priors placed over the mean rate and the default gamma (α=0.5396, β=0.3819)
329 priors being placed over the standard deviation [45]. The bumblebees were
330 constrained to be monophyletic, with the honeybee, *Apis mellifera*, as an outgroup. A
331 gamma (α=74.85889, β=0.4366812) distributed divergence time prior was placed
332 over the tMRCA of the *Bombus* clade, with parameters optimised to match the 2.5th
333 and 97.5th percentiles of the posterior distribution of ages previously estimated by
334 Hines (2008) [46]. Four separate runs of the MCMC were performed for 100,000,000
335 steps from random starting trees, with the first 50,000,000 steps being discarded as
336 burn in. Convergence of the posterior among runs was assessed in Tracer v1.7 [47].
337 The posterior sample was thinned to 1000 trees.
338

339 For the virus phylogeny, amino acid sequences were inferred based on the translated
340 ORFs for regions predicted to contain RdRp motifs using the GenomeNet MOTIF
341 search function [48] against the Pfam database [49], with an expectation cut-off of
342 0.00001. If a virus had no annotated motifs, the canonical GDD RdRp amino acid
343 motif [50] was identified manually. Additional virus species (Supp Table 5) were
344 added to the phylogeny to anchor species with short generated contigs, and to break
345 up long branches. Given the long evolutionary distance between the viruses,
346 PROMALS3D [51] was used to align viral sequences. The alignments were trimmed
347 to the first conserved secondary structural element at both ends as predicted by
348 PROMALS3D with the 0.95 conservation metric. Two of the novel viruses (Agassiz
349 Rock virus and Cnoc Mor virus) were not included in this phylogeny because the
350 section of the RdRp gene required fell outside the available contig. Given that it is
351 unclear whether there was a universal common ancestor of all RNA viruses [52], we
352 aligned the sequences and generated the phylogeny twice, with and without the
353 negative sense RNA viruses (Supp Table 5). The trees serve purely to quantify
354 expected variance (under a Brownian motion model of evolution) between closely
355 related viruses. The deep splits in the phylogeny are poorly resolved with RdRp data
356 [53], due to the fast evolutionary rates of RNA viruses, the considerable time since
357 divergence and permutations in the RdRp sequence [54]. However, this should not
358 overly bias the conclusions as beyond a certain evolutionary distance, the viruses
359 would be expected to become essentially uncorrelated when averaged across the
360 posterior (Supp Table 6 for realised correlations).
361
362 Phylogenetic models used the BLOSUM62 rate matrix [55] with gamma distributed
363 rate variation using 4 gamma categories, an uncorrelated lognormal relaxed clock [45]

364 and a Yule tree prior. A CTMC rate reference prior [56] was placed over the clock
365 mean and an exponential (λ=1) prior was placed over the standard deviation. The
366 alpha parameter of the gamma distributed rate variation was given an exponential
367 (λ=1) prior. Absolute dating of viral trees is difficult due to the inconsistency in
368 estimated ages provided by estimated clock rates and known orthologous insertions
369 between sister host species [57], but is not essential for our analysis, which depends
370 only on relative branch lengths. Nevertheless, we chose to use orthologous insertions
371 (related historical viruses stably integrated into the genomes of species with better
372 known divergence dates) to provide approximate dates for our tree. To account for the
373 estimated ages of RNA viral families [58], we set a uniform lognormal prior with an
374 offset of 97 Mya, a mean of 500 Mya and a logged standard deviation of 0.5 on the
375 age of the root of the tree including the negative sense RNA viruses and a lognormal
376 prior with an offset of 76 Mya, a mean of 500 Mya and a logged standard deviation of
377 0.5 on the age of the tree excluding them. Two partitiviruses (Rosellinia necatrix
378 partitivirus 2 and Raphanus sativus cryptic virus 1) known to have a common ancestor
379 older than 10 Mya [59] were included for dating purposes. We placed a diffuse
380 lognormal prior with an offset of 10 Mya, a mean of 30 Mya and a logged standard
381 deviation of 0.5, on the age of the MRCA of these species. Both models were run over
382 10 separate chains for 50,000,000 generations on a cluster in BEAST v1.8.4 [60],
383 with 25,000,000 generations being discarded as burn-in. Convergence of the posterior
384 was assessed in Tracer v1.7 [47]. The posterior distributions were combined and
385 thinned to 1000 trees.
386
387 **Prevalence Estimation**

388 Maximum likelihood prevalence and 2-log-likelihood confidence intervals were
389 estimated for each host/virus combination with more than one pool using the code
390 from Webster et al. (2015) [61]. As the samples were small pooled groups of
391 individuals, such that a PCR 'positive' represents one or more infections, we
392 modelled the prevalence using a "pooled binomial" likelihood [62–64]. This approach
393 requires that the underlying prevalence of a virus is the same in all pools, which is
394 unlikely for bumblebees sampled from different locations. Estimates should therefore
395 be treated with caution.
396
397 **Co-phylogenetic Mixed Model Analysis**
398 To test for the evolutionary effects on association, the presence/absence data and the
399 phylogenetic trees were analysed using a co-phylogenetic mixed model [25]
400 implemented in Stan [65]. Our model is explicitly focused at the individual-level, and
401 the model's predictions represent the predicted probability of infection within an
402 individual of the species. This is in contrast with Hadfield et al.'s original
403 implementation where the focus was at the species- or population- level and the
404 model was estimating the probability that the parasite would be found in the species
405 or population at all. In all cases, the presented models showed no divergences,
406 acceptable Rhat and E-BFMI values and effective sample sizes of over 200.
407
408 We fitted host and virus phylogenetic effects, which measure the extent to which
409 variation in prevalence is clustered on the host and viral phylogenies respectively. We
410 also fitted host and viral evolutionary interaction effects, which measure the extent to
411 which related species have similar probabilities of infection in the sets of their
412 interaction partners. The final phylogenetic term fitted was a coevolutionary

413 interaction, which measures the extent to which related hosts are infected to similar
414 degrees by related viruses (see Box 1).
415
416 In addition to the phylogenetic terms, non-phylogenetic host and virus terms, an
417 interaction between these terms and a pool ID term were fitted. The non-phylogenetic
418 host and virus terms measure variation that can be partitioned between host species
419 and virus species in average infection risk that is not consistent with trait evolution by
420 Brownian motion. The interaction term measures variation that can be partitioned
421 between pairwise interactions between individual hosts and viruses that is not
422 consistent with the linear sum of their individual means from the non-phylogenetic
423 host and virus terms. The pool ID effect measures variation between pools in
424 infection risk averaged over all the viruses tested. As the pools combined hosts by
425 species rather than by location, so that some had individuals from multiple locations,
426 we treated each location and each realised combination of locations as levels of a
427 random effect, terming this the "spatial composition effect". This describes the
428 variation in average infection level between realised combinations of locations
429 averaged across viruses. Model 1 included all the viruses, Model 2 excluded the
430 negative sense RNA viruses and Model 3 fitted a pseudo-taxonomic model. In Model
431 3, the relationship among the viruses was represented by a polytomic viral tree with
432 arbitrary branch lengths (with a root-to tip distance of 1 unit, and equal length
433 between each taxonomic level) with the viruses being split first by their genomic type
434 (+ve sense RNA, -ve sense RNA and dsRNA) implying a covariance of 0 between
435 genome structures, followed by splitting by the putative viral clades identified by Shi
436 et al. (2016) [31]. This was done to test for potential bias caused the by the possibility

437 of systematic misidentification of the correct relationship between families in the
438 estimated viral trees.
439
440 The form of the models is shown below, where i indexes the data points, group_i
441 represents the level of a categorical variable that the ith pool belongs to, y_i represents
442 the 1/0 indicator for the presence or absence of infection in the ith pool, k_i represents
443 the number of individuals in the ith pool, p_i is the unmeasured probability of infection
444 of a single individual in the ith pool, y'_i is the estimated value of $\log_e(p_i/(1 - p_i))$, μ is
445 the global mean of the latent variable, ε is a normally distributed error term. All terms
446 were fitted as random effects (i.e. estimated by partial pooling). As above, a "pooled
447 binomial" likelihood was used [62–64].
448

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

449 $$p_i = \exp(y'_i)/\exp(1 + y'_i)$$

450 $$y'_i = \mu + \text{host}_i + \text{virus}_i + \text{interaction}_i + \text{host phylogenetic effect}_i + \text{virus}$$
451 $$\text{phylogenetic effect}_i + \text{host evolutionary interaction}_i + \text{virus evolutionary}$$
452 $$\text{interaction}_i + \text{coevolutionary interaction}_i + \text{pool ID}_i + \text{species composition}_i + \varepsilon$$
453
454 All variance-covariance matrices were generated as described in Hadfield et al.
455 (2014) [25], with the variance-covariance matrices scaled to correlation matrices. A
456 standard logistic prior was placed over the global intercept on the latent scale, μ,
457 representing a flat prior on the probability scale. An exponential (λ=1) prior was
458 placed on each variance term in the model. In the full model with all variances being
459 estimated, this is equivalent to a gamma (α=11, β=1) prior over the total variance,
460 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 [66] and 90% credible intervals are reported as these are more robust to sampling in the tails of the posterior distribution [67].

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 shown 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 et al (2005) [68]. 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 therefore its ecology differs from the other species in such a way that tongue length would not be expected to be correlated with the viral community distance.

In order to test for a correlation between tongue length and viral assemblage similarity, estimates of the distance in viral communities between host species are required. These were generated as follows: For each host-virus combination, the

511 package ‘prevalence’ was used to generate posterior draws of the underlying
512 prevalences under a Beta (1,1) prior. Then 1000 draws per species were taken from
513 these sets of MCMC draws to generate 1000 matrices of host-virus prevalences
514 consistent with the raw data. For each of these matrices, the distance between each
515 species’ viral community was calculated by taking the vector of estimated prevalences
516 for the 16 viruses of a given species as a coordinate in a 16-dimensional space then
517 calculating the Euclidean distance between these points. The rank correlation
518 (Kendall’s τ-b) between each pair of species’ viral community distances and their
519 tongue length distances was then calculated, using the mantel function in the R
520 package vegan. The point estimate presented is the median of the 1000 initial
521 correlations accounting for the uncertainty in the underlying prevalences. The 95%
522 confidence interval is the 2.5^{th} and 97.5^{th} percentiles distribution of estimated
523 correlations.
524
525 **Results**
526
527 RNA was extracted from 13 species of bumblebee from nine sites, to identify new
528 viruses, assay their prevalence and their pattern of distribution across host species and
529 to test whether the evolutionary histories of the viruses and hosts have impacted the
530 current distribution.
531
532 **Read and Assembly Statistics**
533 A total of 134,026,056 sequencing read pairs were generated for *Bombus lucorum*,
534 135,590,922 for *Bombus terrestris*, 128,670,194 for *Bombus pascuorum* and
535 26,838,390 for the other *Bombus* species with 0.37, 0.38, 3.36 and 15.12 percent of

536 reads mapping to the known viruses or the novel bee viruses found in the study. The
537 poly-A and DSN normalized datasets were unexpectedly highly correlated, given their
538 expected biases (1.000 for *Bombus terrestris*, 0.999 for *Bombus pascuorum* and 0.998
539 for *Bombus lucorum*) implying that the sequencing results were highly consistent
540 irrespective of the selection method used.
541
542 **Previously Described Viruses Present in the Metagenomic Pools**
543 RNAseq reads mapped to three previously described bee viruses. The majority of
544 these reads mapped either to the Acute bee paralysis virus/Kashmir bee virus complex
545 (henceforth ABPV) [29] or to Slow bee paralysis virus (SBPV) [28]. Additionally, in
546 the mixed *Bombus* pool, reads were found mapping to Hubei partiti-like virus 34
547 (HPLV34) a virus initially detected, though not named, in honeybees by Cornman et
548 al (2012) [30], then subsequently also reported in a sample from Chinese landsnails
549 by Shi et al (2016) [31].
550
551 No RNAseq reads were mapped to Deformed wing virus – type A [69], Chronic bee
552 paralysis virus [29], Bee macula-like virus [70], Ganda bee virus [71], Scaldis River
553 bee virus [71], Black queen cell virus [72], Apis rhabdovirus 1 [73], Apis rhabdovirus
554 2 [73], Apis bunyavirus 1 [73], Apis bunyavirus 2 [73], Apis flavivirus [73], Apis
555 dicistrovirus [73], Apis Nora virus [73] and members of the Lake Sinai virus complex
556 [74]. A small number of small RNA reads did map to these viruses, however, this
557 likely represents cross-mapping, given the lower stringency of 22nt reads. Two of the
558 viral contigs generated by the *de novo* assembly had high similarity to previously
559 described plant viruses; both RNAs of White clover cryptic virus 2 [75] (96%
560 identity), both RNAs of strain of Arabis mosaic virus (MH614320/MH614321) [76]

561 distant to previously sequenced strains (91% identity) and a strain of Red Clover
562 nepovirus A (MH614312) [77] distant to previously sequenced strains (90% identity).
563
564 **Putative Novel Viral-like Sequences**
565 We identified 37 putative novel viral contigs, four mapping to DNA viruses (4
566 densovirus-like contigs) and 33 to RNA viruses (4 Reo group contigs, 2 Toti-Chryso
567 group contigs, 4 Bunya-Arena group contigs, 1 Orthomyxoviridae-like contig, 8
568 Hepe-Virga group contigs, 12 Picorna-Calici group contigs and 2 Tombus-Noda
569 group contigs). Based on the supposition that a contig represents a separate virus if it
570 maps to a different viral grouping than the other contigs, or if it can be aligned to all
571 other contigs within its assigned viral grouping, this represents 30 new viruses with
572 seven remaining contigs that may represent other genomic regions of these 30 viruses
573 or separate viruses that cannot be confirmed as such. See Table 1 for information on
574 the viruses tested for prevalence using PCR and Supplementary Table 7 for detailed
575 information on all of the identified contigs. The numbers of reads mapping these
576 contigs were variable and are shown in Table 2.
577
578 **Table 1** The names, genome structures and groupings (following Shi et al. (2016) [31]) of the newly
579 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

580

581 **Table 2** The RNAseq reads per kilobase per mapped million reads in the *Bombus terrestris*, *Bombus*
582 *lucorum, Bombus* pascuorum and mixed *Bombus* pools. Structural zeros are indicated by dashes, zeros
583 in the table indicate below 0.005. Contigs with names in bold meet the criterion of having at least 50
584 mapping small RNA reads with a sharp peak in the size distribution at 22nt in *Bombus terrestris,*
585 *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

586

587 **siRNA-based Evidence for Infection**

588 RNA interference is an important component of antiviral defence in arthropods [78].
589 As part of this defence mechanism, homologs of *Drosophila* Dicer-2 cleave dsRNA,
590 usually in the form of replication intermediates, giving rise to a characteristically
591 narrow and sharply peaked distribution of virus-derived small RNAs. Thus the
592 presence of such small RNAs from both strands of an ssRNA virus provide
593 compelling evidence that the virus was replicating. In bumblebees the characteristic
594 Dicer-mediated viral siRNAs peak sharply at 22nt [73], and the viruses that displayed
595 at least 50 characteristic viral siRNAs are marked in Table 2. The distribution of the
596 mapped small RNA reads is shown in Fig 3 for all viruses where the siRNAs are
597 described in the main text, with full data in Supp Fig 1.

598

599 In the three bumblebee species with siRNA data, a sequence similar to a phlebovirus
600 glycoprotein (AEL29653.1) displayed >50 siRNA reads. However, the size spectra of
601 these reads is centered on 24nt with a strong bias for a 5' terminal uracil, with the
602 antisense mapping orientation being more prevalent. This 5' U-bias is consistent with
603 insect piRNAs [79], and the predominant antisense orientation is consistent with the

604 piRNA mapping pattern to endogenous viral elements (EVE) in mosquitoes [80].
605 However, the size of piRNAs in bumblebees is generally larger than this [81]. This
606 sequence is therefore potentially an EVE that has either been gained multiple times or
607 has been maintained in the bumblebee genome since at least the *B. pascuorum*-*B.*
608 *terrestris*/*B. lucorum* split.
609
610 It is notable that the size distribution of viral siRNAs is less sharply peaked in
611 Mayfield virus 1, Mayfield virus 2 and Slow bee paralysis virus (excepting Mayfield
612 virus 1 in *B. lucorum*, which is sharply peaked), with broad 'shoulders'. This is
613 reminiscent of the pattern seen for Drosophila C virus and Drosophila Nora virus in
614 wild-collected D. *melanogaster* [61], both of which contain a viral suppressor of
615 RNAi [82,83].
616
617 *B. pascuorum* also had siRNA reads mapping to a sequence with 90% identity to Red
618 clover nepovirus A (MG253829.1) [77]. However, the read length spectra were
619 sharply peaked at 21nt, rather than the 22nt of bumblebee viRNAs. This is consistent
620 with siRNA's produced from DLC4, the key antiviral dicer in *Arabidopsis thaliana*
621 [84] implying acquisition of the small RNAs through nectar or pollen contamination.
622

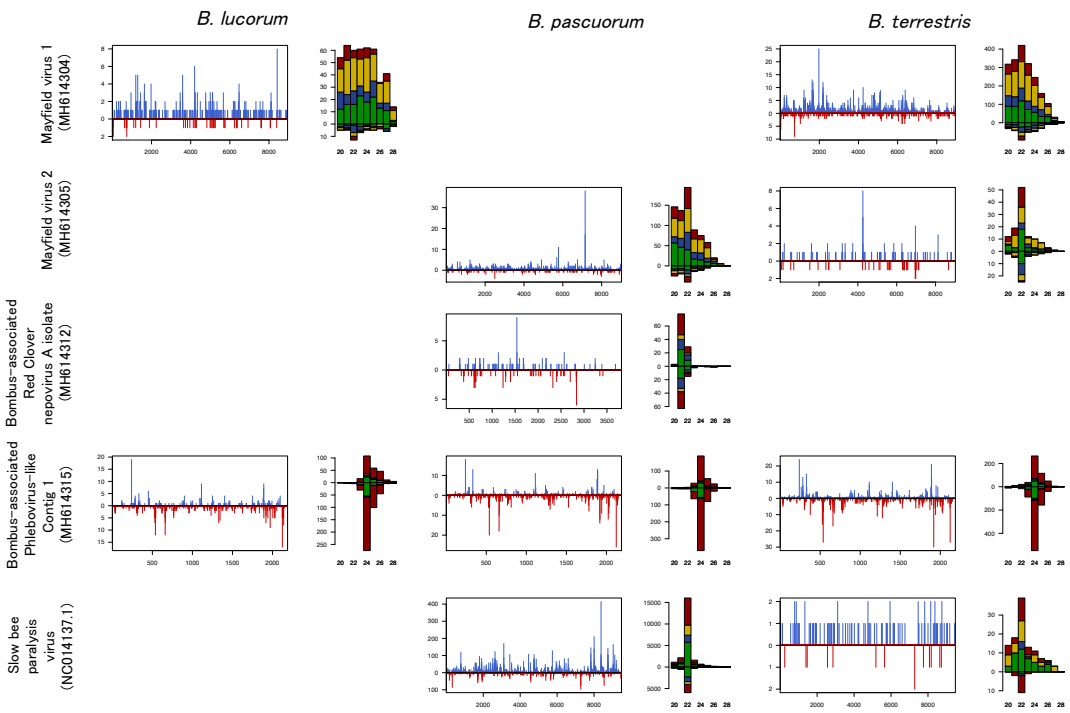
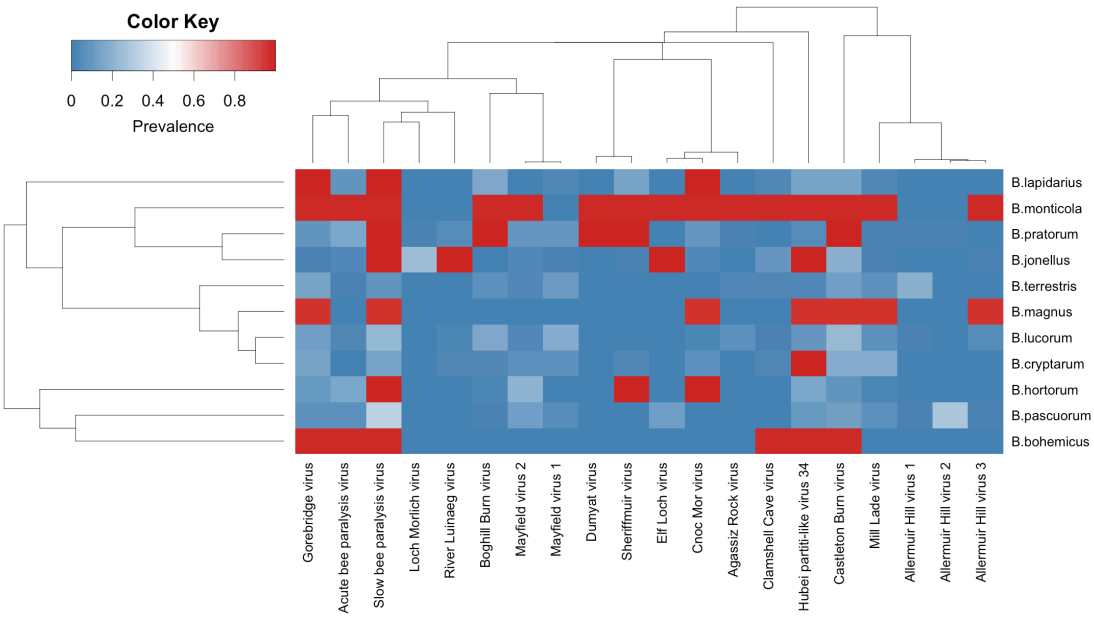


Fig 2 The mapping of small RNA reads to Mayfield virus 1, Mayfield virus 2, a contig similar to Tomato black ring virus, a contig similar to a phlebovirus glycoprotein and Slow bee paralysis virus. Blue lines represent reads mapping to the positive-sense strand at that genomic position, red lines represent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

Prevalence

Species level prevalences differed dramatically among the different viruses (Fig 3). Prevalences were generally low to intermediate, with modal viral prevalences for most host-virus combinations being below 15%. Slow bee paralysis virus was by far the most common virus in the sample, with estimated prevalences of greater than 25% in multiple species. Our ability to estimate the prevalence of common viruses is limited by the pooling, leading us to only be able to assign lower bounds to

639 prevalences in these cases, but in 7 of 11 species, all pools were positive for SBPV.
640 Acute bee paralysis virus, Hubei partiti-like virus 34, Castleton Burn virus,
641 Gorebridge virus, Mayfield virus 1 and Mayfield virus 2 all reached 15-25%
642 prevalences in multiple species. Several viruses showed strong signals of species
643 specificity, having very low to zero prevalences in multiple host species but high
644 prevalences in others. Examples of this pattern include Allermuir Hill virus 1 in *B.*
645 *terrestris,* Allermuir Hill virus 2 in *B. pascuorum*, Allermuir Hill virus 3 in *B. magnus*
646 and *B. monticola*, as well as Loch Morlich virus and River Luinaeg virus in *B.*
647 *jonellus*.
648



649

650 **Fig 3** A heatmap of maximum likelihood estimates for prevalence. Hosts and viruses are ordered by
651 phylogenetic relatedness, the trees represent the maximum clade credibility topology. Squares in red
652 with maximum likelihood estimates of the prevalence of 1 correspond to cases where all pools were
653 positive. The maximum likelihood estimate is likely extremely upwardly biased in this case.
654

655 **Host-Pathogen Co-phylogenetic Models**

656 All models that included a virus phylogeny term gave qualitatively similar results (Fig

657 4, Table 3). This suggests that the results are robust to both phylogenetic uncertainty
658 and the assumption of a common ancestor of all RNA viruses. For this reason, for the
659 rest of this section, estimates will be given from the model containing the estimated
660 phylogeny with all the RNA viruses included. All estimates represent the percentage
661 of the total variance in the model (the sum of all estimated variance components
662 adjusted for the variance of the link function by the addition of $\pi^2/3$) explained by a
663 term. In all cases, the presented point estimate is the posterior mean, and 90% shortest
664 posterior density intervals [66] are presented following in square brackets. Shortest
665 posterior density intervals are a variant of highest posterior density intervals and
666 describe the shortest possible interval containing (in this case) 90% of the probability
667 density for the parameter. We present 90% intervals rather than the standard 95%
668 intervals, as 95% intervals calculated from simulation draws are less computationally
669 stable [67]. In all cases but the virus phylogenetic effect, the posterior estimates for
670 the proportion of variance explained by each effect differed strongly from their
671 induced priors (Supp Fig 2).
672
673 **Summary of Model Results**
674 We find evidence that which viral species infects a host, the specific interaction
675 between individual hosts and individual viruses and related hosts having similar
676 prevalences with the same sets of viruses all explain variation in infection prevalence.
677
678 **Total Evolutionarily-associated Variation**
679 In the models containing the virus phylogeny, approximately a quarter (25.9% [11.6-
680 40.4]) of the total variation in prevalence was explained by terms accounting for the
681 evolutionary histories of hosts and viruses (host phylogenetic effect, virus

682 phylogenetic effect, host evolutionary interaction effect, virus evolutionary interaction
683 effect and coevolutionary interaction).
684
685 **Host and Virus Level Effects**
686 The host and virus phylogenetic effects measure the extent to which related hosts
687 have similar average prevalences of virus infection and related viruses have similar
688 average prevalences across hosts. The host and virus non-phylogenetic terms measure
689 the extent to which hosts and viruses differ in their average infection levels in manner
690 not consistent with evolution by Brownian motion along a phylogeny. The host
691 species and phylogenetic effects explained a small proportion of the total variance in
692 infection probability (species: 1.9%, [0.0-4.6]; phylogenetic: 2.9%, [0.0-6.5]). The
693 shape of the posterior distributions for the two parameters visualised in Fig 4, makes
694 it clear that the most credible values for both of these parameters are 0. While it is
695 unlikely that there is no variation in average prevalence between host species, it is
696 clear that the amount of prevalence explained by hosts differing in their infection
697 levels averaged across viruses is small relative to the other effects.
698
699 The virus species and phylogenetic effects explained a larger proportion of the total
700 variance in infection probability, with non-phylogenetic variation dominating, but
701 were imprecisely estimated (species: 13.8%, [3.8-23.7]; phylogenetic: 8.5%, [0.0-
702 17.8]). The posterior density for the phylogenetic effect is concentrated at 0. So, while
703 it is clear there is a virus species effect, the data does not appear informative for the
704 presence or absence of a viral phylogenetic effect in this system. The posterior draws
705 for the viral species and phylogenetic effect were negatively correlated within
706 iterations, indicating that the model had difficulty partitioning the two. This partial

707 non-identifiablility explains the broad posteriors on both.

708

709 **Interaction Effects**

710 A host evolutionary interaction effect measures the extent to which more closely
711 related hosts have more similar prevalences with the same sets of viruses, and a virus
712 evolutionary interaction effect measures the extent to which more closely related
713 viruses infect the same sets of hosts to more similar degrees. A coevolutionary
714 interaction term measures the extent to which more related hosts are infected to
715 similar degrees with viruses that are themselves related. The non-phylogenetic
716 interaction term measures the extent to which there is variance in the mean prevalence
717 of specific host-virus pairings, that are is not consistent with the other interaction
718 effects (see Box 1).

719

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

724

725 There was evidence for a host evolutionary interaction explaining some of the total
726 variance in prevalence (9.5% [0.0-18.3]). This is the only parameter in the model for
727 which the estimated size of the effect depended strongly on the specific treatment of
728 the virus phylogeny (see Fig 4). Whether the lower 90% bound of the credible interval
729 for the variation explained rounded to 0.0 or 0.1 depended on the phylogenetic matrix
730 (or set of phylogenetic matrices) inputted. The marginal posterior distribution of the
731 parameter was concentrated at lower values when the estimated phylogeny including

732 the negative-sense RNA viruses was used, and at higher values in the other two cases.
733 However, irrespective of the choice of virus phylogeny, the mode of the distribution
734 and majority of the density was distant from zero, implying that the effect is likely to
735 be biologically relevant. As the virus phylogenies themselves are not actually directly
736 involved in this term, this must be due to the partitioning of variance across other
737 terms being cryptically different depending on the assumptions about the virus
738 phylogeny.
739
740 There was also a clear non-phylogenetic interaction (15.8% [5.8-25.4]), implying that
741 much of the variation in prevalence is due to the idiosyncrasies of individual host-
742 virus combinations.
743
744 As with the virus species effect and the virus phylogenetic effect, the MCMC draws
745 for the proportion of variance explained by the host evolutionary interaction and non-
746 phylogenetic interaction were negatively correlated within an iteration, implying that
747 separating these parameters was proving difficult. While this led to a diffuse posterior
748 with wide credible intervals for both, they remain individually interpretable, and both
749 effects appear present simultaneously.

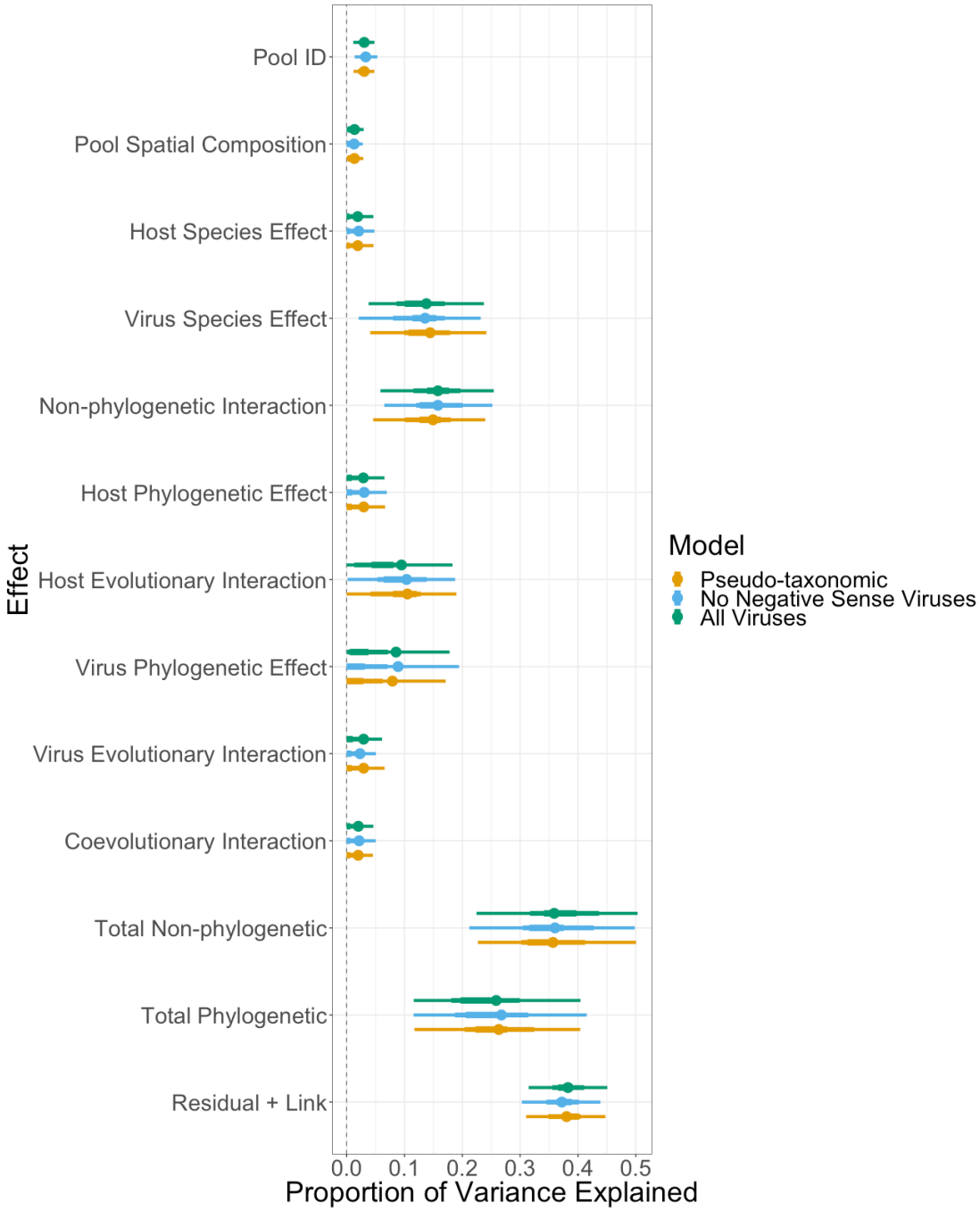


Fig 4 Comparison of estimated proportion of variance in prevalence explained by different parameters between models. As each factor explains a proportion of the attributed variance in the model, total variance over all factors must sum to 1. For each parameter, the circle represents the modal estimate, 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

759 proportion of variation in prevalence explained by hosts having different average viral prevalences.
760 "Virus Species Effect" is the proportion of variation in prevalence explained by viruses differing in
761 their average prevalences. "Non-phylogenetic Interaction" is the proportion of variation in prevalence
762 explained by host-virus combinations differing in the their average prevalences beyond that which
763 would be expected by their host and virus species effects alone. "Host Phylogenetic Effect" is the
764 proportion of variation in prevalence explained by hosts having average viral prevalences correlated
765 across the host phylogeny. "Virus Phylogenetic Effect" is the proportion of variation in prevalence
766 explained by viruses having average prevalences correlated across the viral phylogeny. "Host
767 Evolutionary Interaction" is the proportion of variation in prevalence explained by related hosts having
768 correlated viral assemblages. "Virus Evolutionary Interaction" is the proportion of the variation
769 explained by related viruses having correlated host assemblages. "Coevolutionary Interaction" is the
770 proportion of the variation explained by related hosts having similar prevalences of related viruses.
771 "Total Non-phylogenetic" is the proportion of the variation that can be explained by terms not
772 involving the host and virus phylogeny and excluding the residual ("Host Species Effect", "Virus
773 Species Effect", "Pool ID", "Spatial Composition Effect", "Non-phylogenetic Interaction"). "Total
774 phylogenetic" is the proportion of the variation that can be explained by terms involving a host or virus
775 phylogeny ("Host Phylogenetic Effect", "Virus Phylogenetic Effect", "Host Evolutionary Interaction",
776 "Virus Evolutionary Interaction", "Coevolutionary Interaction"). "Residual + Link" is the proportion of
777 the total variance that is explained by the residual variance and variance of the logistic distribution
778 ($\pi^2/3$).
779
780

781 **Table 3** Mean estimates for the intra-class correlations of each variance component. The point estimate
782 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)

783

784

785 **Tongue Length-Viral Community Correlation**

786 Given that the co-phylogenetic model found that related hosts share viral
787 communities, one potential mechanism for this is phylogenetically-biased exposure,
788 driven by phylogenetically correlated floral preferences. If bumblebee species with
789 similar flower preferences had similar viral communities, it would be expected that
790 there would be a positive correlation between tongue length similarity (as this is an
791 important factor in floral preference) and viral community similarity between pairs of
792 species. The point estimate of the correlation between the two distances was small
793 and negative (-0.06), but the 95% confidence intervals for that point estimate
794 overlapped zero (-0.13, 0.00), so given the uncertainty in the data, a correlation of
795 zero cannot be rejected. Nonetheless, given this data, a strong positive relationship
796 between tongue length and viral community similarity seems unlikely, a result

797 inconsistent with phylogenetically-biased exposure driven by tongue length-mediated
798 floral choice.
799
800 **Discussion**
801 Using wild bumblebee species that share transmission opportunities, we have shown
802 that variation in the prevalence of infection in the wild is explained by related hosts
803 being infected with the same viruses to similar degrees, viruses differing in their
804 average prevalence and individual virus-host pairings having greater or lesser
805 prevalence than would be expected by the host and virus species effects alone.
806
807 **Virus Discovery**
808 There is now an extensive diversity of viruses known in bees, with most new studies
809 finding novel viruses [30,71,73,74,85-87]. We have found 37 novel putative viral
810 contigs in the transcriptomes of wild-caught bumblebees from across Scotland,
811 suggesting that virus discovery in this taxonomic group is far from saturation. As with
812 any metagenomic study, it is hard to be confident that the virus-like contigs represent
813 real infections of the sampled host, rather than surface or gut contaminants. However,
814 the presence of 22nt virus siRNAs, generated from double-stranded viruses by Dicer
815 as part of an antiviral response in the host, provides compelling evidence that at least
816 10 of these contigs (Densovirus 2 and 3, Elf Loch virus, Allermuir Hill virus 1, 2 and
817 3, Mill Lade virus, Mayfield virus 1 and 2, and Castleton Burn virus) represent active
818 viral infections in bumblebees.
819
820 Mites and nematodes both parasitise bumblebees and therefore could potentially be an
821 alternative source of the small RNAs. Mite viRNAs are reported to be centered at

822 24nt [73], and could therefore not produce the small RNA patterns observed.
823 Nematode viRNAs are centered at 22nt, like bumblebee viRNAs, [88] and thus could
824 potentially produce this pattern. While, outside of queens infected with *Sphaerularia*
825 *bombi*, nematode infection of wild bumblebees appears to be very rare [89],
826 nematodes cannot be categorically ruled out as a source of the observed small RNAs.
827 One contig's (MH614312) mapped small RNAs were centered at 21nt, and the closest
828 known virus was a nepovirus of plants. As DCL4, a major plant Dicer, produces
829 viRNAs of this size [90], this is consistent with that particular virus being a plant
830 virus, which was transferred in collected nectar or pollen.
831
832 **Phylogenetic Effects**
833 We found no evidence for a large host phylogenetic effect, where related hosts have
834 correlated average viral prevalences. To the best of our knowledge no studies have
835 previously applied these methods to viruses sampled from wild animals. However,
836 other traits relating to viral disease in a series of studies in *Drosophila* species under
837 experimental conditions have consistently detected host phylogenetic effects in
838 factors that would be expected to be correlated with prevalence in the wild, such as
839 infection probability [24], virulence and viral load [91] and viral load alone [92].
840 However, two of these studies focused on a single isolate of Drosophila C virus.
841 Therefore, the variation that they attribute to a phylogenetic effect may be partitioned
842 into the host evolutionary interaction in our study, as a host evolutionary interaction is
843 equivalent to separate inconsistent host phylogenetic effects for each virus. Our data
844 were not particularly informative for the presence or absence of a virus phylogenetic
845 effect, with the posterior being very diffuse with a majority of the density near zero.
846 This appears partly due to difficulties partitioning the variation between the virus

847 species effect and virus phylogenetic effect. Irrespective of the cause, we can make no
848 strong statements about whether related viruses exhibit similar prevalences across
849 hosts from this dataset.
850
851 **Host Species Effect**
852 There was little evidence for an important effect of host species independent of the
853 phylogeny, implying that hosts do not strongly vary in the average degree to which
854 they are infected with viruses. The previous studies using these methods have
855 universally found host species effects [25,93]. However, as both of these studies have
856 used mammal-eukaryotic parasite datasets, the degree of relevance for them as a
857 comparison in unclear. Experimental evidence from virus studies across drosophilid
858 flies have found weak to zero host species effects on the titre of sigma viruses [24]
859 and Drosophila C virus virulence [91] but considerably larger host species effects on
860 Drosophila C virus load [91]. This between-study variation potentially indicates a
861 reason we did not detect a host species effect. With a single pathogen, the average and
862 particular degrees of variation in infection between hosts are identical. As soon as
863 multiple pathogens are involved, they diverge, such that it is possible for there to be
864 no variation in the average prevalence between hosts, but still considerable variation
865 in the prevalence of particular viruses between hosts, which is consistent with the
866 presence of a host species effect in correlates of prevalence in some viruses but not
867 others, as noted above.
868
869 **Virus Species Effect**
870 A clear effect of virus species independent of the virus phylogeny was detected in the
871 dataset, despite the uncertainty added by the difficulty partitioning the virus species

872 and phylogenetic effects. Therefore, viruses differed in their prevalences averaged
873 across hosts. This is an intuitive result as viruses differ in host range [94], virulence
874 [95,96] and infectious period [96] at both the species and the strain level. Variation in
875 host range changes the size of the host pool available for infection, and variation in
876 virulence and infectious period both change the length of time any infected host is
877 available for sampling. All these factors would be expected to drive consistent
878 differences in long-run prevalences between viruses. Additionally, as our sites were
879 only sampled once, short-term effects will also drive between virus variation. Any
880 virus that was experiencing an epizootic at the time of sampling will be
881 overrepresented relative to its long-run prevalence, further increasing the between
882 virus variation.
883
884 **Host Evolutionary Interaction Effect**
885 A host assemblage effect was found, where phylogenetically related hosts share viral
886 assemblages, showing that more closely related hosts are more similar in virus
887 prevalence for groups of viruses. The statistical machinery required for estimating this
888 effect is quite new and, in the disease ecology field, has predominantly been applied
889 to mammal-parasite and plant-parasite systems where there are good datasets already.
890 Nonetheless, host evolutionary interactions have always been found when searched
891 for using these methods [25,93] and analogous effects are commonly found using
892 different methods [97–99]. In a system where these viruses were host limited, this
893 pattern could be explained by preferential host shifting, where parasites more
894 frequently gain the ability to infect hosts closely related to ones that they are already
895 capable of infecting. Preferential host shifting is known to be a general phenomenon,
896 and has been observed in macroparasites, viruses and protozoans (see Longdon et al

897 (2014) [27] and the references within).

898

899 While some of the viruses in this study were not detected in a subset of host species,
900 most of the viruses found here appear to genuinely be multihost viruses, with the
901 majority being detected in over half the sampled species. Given this, a combination of
902 biased cross-species transmission and preferential host shifting appears a better
903 explanation in this system. Biased cross-species transmission occurs when
904 transmission occurs more frequently between some species that a pathogen is already
905 capable of infecting than others. This biased cross-species transmission could be
906 driven by two non-exclusive mechanisms: phylogenetically-biased transmission
907 probabilities and phylogenetically-biased exposure.

908

909 Phylogenetically-biased transmission probabilities occur when cross-species
910 transmission is more frequent among close relatives, due to the probability of
911 infection after contact with the virus being similar between related species. Related
912 hosts present correlated environments from the perspective of the virus at the
913 molecular and anatomical level, therefore adaptation to one should provide
914 corresponding fitness increases on the other. Experimental results have shown that
915 correlated mutations occur on viral entry to related hosts [100], implying that this
916 cross-adaptation does occur. However, this is probabilistic, and the mutations fixed on
917 entry can differ between replicate entries [100,101]. Therefore, if there is antagonistic
918 pleiotropy between mutations that are adaptive in two different groups of hosts and
919 cross-adaptation predicts the probability of successful infection on contact, then a
920 phylogenetically-biased transmission network will result.

921

922 Phylogenetically-biased exposure represents an evolutionarily-driven ecological
923 phenomenon that biases cross-species transmission rates, mediated by niche overlap.
924 Contaminated flowers are likely to be an important source of intra- and inter- specific
925 pathogen transmission in bumblebees and pollinators more generally [102–104]. The
926 flower visitation network has been shown to be associated with the partitioning of
927 genetic diversity of *Crithidia bombi* between bumblebee hosts [9], and the network
928 itself is highly structured, though temporally variable [105]. Different bumblebee
929 species show tongue length differences, which are phylogenetically associated [15],
930 and the differences in tongue length correlate with differential flower usage between
931 bumblebee species [12,106]. If infection occurs at contaminated flowers, the
932 structuring of the flower usage network could cause different flowers to build up
933 different surface viral communities. This could drive consistent phylogenetically-
934 correlated differences in viral infection rates through differential exposure.
935
936 Post-hoc testing did not find a positive relationship between tongue length
937 dissimilarity (a rough proxy for species-level flower choice dissimilarity) and viral
938 community composition dissimilarity, which provides some evidence against
939 phylogenetically-biased exposure as the causative mechanism. However, the study
940 design in this case is not optimal for disentangling biased transmission probabilities
941 and biased exposure, as species were sampled from different locations at a single
942 timepoint and prevalence of the viruses varied spatially. Given this, drawing strong
943 conclusions as to the relative impact of the two mechanisms outlined above based on
944 this data would be premature. Similarly, the subgenus *Psithyrus* contains socially
945 parasitic species that are coevolved to parasitise particular social bumblebee species,
946 which could also lead to phylogenetically-correlated differences in viral infection

947 rates. We were unable to test whether socially parasitic cuckoo bumblebee species
948 have similar viral communities to their hosts, as our study included only a single
949 parasitic bumblebee species, *B. bohemicus*, but the possibility of brood parasitism
950 being an important driver of between-colony disease transmission is worth further
951 study.
952
953 **Virus Evolutionary Interaction Effect and Coevolutionary Interaction**
954 We found no evidence of a large virus evolutionary interaction or coevolutionary
955 interaction. This is largely unsurprising as it would appear implausible that the host
956 assemblages have been conserved over evolutionary time, as the deep splits in the
957 viral families predate the most recent common ancestor of bumblebees by many
958 millions of years [107].
959
960 **Non-phylogenetic Interaction**
961 A non-phylogenetic interaction was detected. This interaction represents variation in
962 prevalence caused by specific host-virus pairings having prevalences beyond that
963 which would be expected by the simple addition of the individual host and virus
964 means. A non-phylogenetic interaction could be caused by a large range of factors,
965 some biological and some due to the specifics of the model, many of which would be
966 likely to be acting simultaneously to generate this signal. One possibility is
967 coevolution between the host and virus that occurred after both diverged from their
968 common ancestor with the closest related species in the study. Another is the
969 complete absence of coevolution, where spillover from a primary or group of primary
970 hosts causes either a constant very low prevalence of dead-end infections, which are
971 none-the-less detectable by PCR. Related to this is a statistical issue involving cases

where not every species in the study is within a virus' host range, and the species that are within the host range are not closely related. In this case, the variation is not absorbed by the host evolutionary interaction and almost no host has a prevalence close to the mean across hosts, as in many species the prevalence is zero, which causes the mean to be considerably lower than the average prevalence in the species the virus does infect. This effect would be magnified if the sampling occurs during an epizootic. More broadly, anything that changes the epidemiological parameters of a virus in a specific host will lead to a non-phylogenetic interaction. Considering the variation in the natural history of viruses and the lesser, but still significant, variation in the natural history of bumblebees, a large non-phylogenetic interaction is to be expected.

Conclusion

While it is clear that viruses are abundant in pollinators, the factors that determine the distribution of pollinator viruses have remained uncertain, outside of a few well-studied cases [108,109]. With the novel viruses discovered in this study, we have investigated predictors of these virus/host associations and found that both the host evolutionary history and the identity of the virus contributes to this distribution. This supports both theory and prior empirical evidence that related species are more at risk of infection from each other's diseases than the diseases of distantly related species. However, the importance of the viral identity and unique interactions between host-virus pairs suggests that the introduction of a novel virus into a community is likely to have unpredictable effects even when no close relatives of currently known hosts are present. This highlights the risk posed by disease spillover for the conservation not

996 only of wild pollinator communities, but also to communities consisting of related
997 animal or plant species in general.
998
999 **Data Availability**
1000 The data and code for running the analyses is available on github under a GPLv3
1001 licence, as code uses code taken from other GPLv3 licenced works
1002 (https://github.com/dpascall/bumblebee-virus-cophylo).
1003
1004 **Acknowledgements**
1005 We thank Jarrod Hadfield for extensive statistical advice, modifications to
1006 MCMCglmm and for helpful comments on this manuscript, Dave Goulson for
1007 assistance identifying some specimens and to Ben Longdon and Bethany Clark for
1008 further comments. Rowan Doff assisted in the lab with RFLP analyses and Claire
1009 Webster with DNAse treatments. Claire Webster, Jarrod Hadfield and Florian Bayer
1010 helped with fieldwork.
1011
1012 **Funding**
1013 This work was funded by a BBSRC SWBIO DTP PhD stipend to DP, a Royal Society
1014 Dorothy Hodgkin Fellowship to LW, Wellcome Trust Research Career Development
1015 Fellowship WT085064 to DJO.
1016
1017 **References**
1018 1. Woolhouse ME, Gowtage-Sequeria S. Host range and emerging and
1019 reemerging pathogens. Emerg Infect Dis. 2005;11(12): 1842–7.
1020 2. Suttle CA. Marine viruses--major players in the global ecosystem. Nat Rev

1021 Microbiol. 2007;5(10): 801–12.
1022 3. Williams PH, Osborne JL. Bumblebee vulnerability and conservation world-
1023 wide. Apidologie. 2009;40(3): 367–87
1024 4. Garibaldi LA, Steffan-Dewenter I, Winfree R, Aizen MA, Bommarco R,
1025 Cunningham SA, et al. Wild pollinators enhance fruit set of crops regardless of
1026 honey bee abundance. Science. 2013;340(6127): 1608–11.
1027 5. Vanbergen AJ, the Insect Pollinators Initiative. Threats to an ecosystem
1028 service: Pressures on pollinators, Front Ecol Environ. 2013;11(5): 251–259.
1029 6. Manley R, Boots M, Wilfert L. Emerging viral disease risk to pollinating
1030 insects: Ecological, evolutionary and anthropogenic factors. J Appl Ecol.
1031 2015;52(2): 331–40.
1032 7. Fenton A, Streicker DG, Petchey OL, Pedersen AB. Are all hosts created
1033 equal? Partitioning host species contributions to parasite persistence in
1034 multihost communities. Am Nat. 2015;186(5): 610–22.
1035 8. Heesterbeek JA. A brief history of R0 and a recipe for its calculation. Acta
1036 Biotheor. 2002;50: 189–204.
1037 9. Salathé RM, Schmid-Hempel P. The genotypic structure of a multi-host
1038 bumblebee parasite suggests a role for ecological niche overlap. PLoS One.
1039 2011;6(8): e22054.
1040 10. Arbulo, N, Santos, E, Salvarrey, S, & Invernizzi C. Proboscis length and
1041 resource utilization in two ruguayan bumblebees: *Bombus atratus* Franklin and
1042 *Bombus bellicosus* Smith (Hymenoptera: Apidae). Neotrop Entomol.
1043 2011;40(4): 483–8.
1044 11. Goulson D, Darvill B. Niche overlap and diet breadth in bumblebees; are rare
1045 species more specialized in their choice of flowers? Apidologie. 2004;35(1):

1046 55–63.
1047 12. Goulson D, Lye GC, Darvill B. Diet breadth, coexistence and rarity in
1048 bumblebees. Biodivers Conserv. 2008;17(13): 3269–88
1049 13. Harder LD. Morphology as a predictor of flower choice by bumble bees.
1050 Ecology. 1985;66(1): 198–210.
1051 14. Lye GC, Kaden JC, Park KJ, Goulson D. Forage use and niche partitioning by
1052 non-native bumblebees in New Zealand: Implications for the conservation of
1053 their populations of origin. J Insect Conserv. 2010;14(6): 607–15.
1054 15. Harmon-Threatt AN, Ackerly DD. Filtering across spatial scales: Phylogeny,
1055 biogeography and community structure in bumble bees. PLoS One. 2013;8(3):
1056 e60446.
1057 16. Goodwin SG. Seasonal phenology and abundance of early-, mid- and long-
1058 season bumble bees in southern England. J Apic Res. 1995;34(2): 79–87.
1059 17. van Riper III C, van Riper SG, Goff ML, Laird M. The epizootiology and
1060 ecological significance of malaria in Hawaiian land birds. Ecol Monogr.
1061 1986;56(4): 327–44.
1062 18. Hueffer K, Parker JS, Weichert WS, Geisel RE, Sgro JY, Parrish CR. The
1063 natural host range shift and subsequent evolution of canine parvovirus resulted
1064 from virus-specific binding to the canine transferrin receptor. J Virol.
1065 2003;77(3): 1718–26.
1066 19. Johnson MB, Lafferty KD, van Oosterhout C, Cable J. Parasite transmission in
1067 social interacting hosts: Monogenean epidemics in guppies. PLoS One.
1068 2011;6(8): e22634.
1069 20. Arneberg P, Skorping A, Grenfell B, Read AF. Host densities as determinants
1070 of abundance in parasite communities. Proc R Soc B Biol Sci. 1998;265(1403):

1283–9.

21. Civitello DJ, Cohen J, Fatima H, Halstead NT, Liriano J, Mcmahon TA, et al. Biodiversity inhibits parasites: Broad evidence for the dilution effect. Proc Natl Acad Sci U S A. 2015;112(28): 8667–71.

22. Curtis VA. Infection-avoidance behaviour in humans and other animals. Trends Immunol. 2014;35(10): 457–64.

23. Chandra RK. Nutrition, immunity, and infection: present knowledge and future directions. Lancet. 1983;321(8326): 688–91.

24. Longdon B, Hadfield JD, Webster CL, Obbard DJ, Jiggins FM. Host phylogeny determines viral persistence and replication in novel hosts. PLoS Pathog. 2011;7(9): e1002260.

25. Hadfield JD, Krasnov BR, Poulin R, Nakagawa S. A tale of two phylogenies: Comparative analyses of ecological interactions. Am Nat. 2014;183(2): 174–87.

26. Rafferty NE, Ives AR. Phylogenetic trait-based analyses of ecological networks. Ecology. 2013;94(10): 2321–33.

27. Longdon B, Brockhurst MA, Russell CA, Welch JJ, Jiggins FM. The evolution and genetics of virus host shifts. PLoS Pathog. 2014;10(11): e1004395.

28. Bailey L, Woods RD. Three previously undescribed viruses from the honey bee. J Gen Virol. 1974;25(2): 175–86.

29. Bailey L, Gibbs AJ, Woods RD. Two viruses from adult honey bees (*Apis mellifera* Linnaeus). Virology. 1963;21(3): 390–5.

30. Cornman RS, Tarpy DR, Chen Y, Jeffreys L, Lopez D, Pettis JS, et al. Pathogen webs in collapsing honey bee colonies. PLoS One. 2012;7(8): e43562.

1096 31. Shi M, Lin XD, Tian JH, Chen LJ, Chen X, Li CX, et al. Redefining the
1097 invertebrate RNA virosphere. Nature. 2016;540(7634): 539-543.
1098 32. Murray TE, Fitzpatrick Ú, Brown MJF, Paxton RJ. Cryptic species diversity in
1099 a widespread bumble bee complex revealed using mitochondrial DNA RFLPs.
1100 Conserv Genet. 2008;9(3): 653–666.
1101 33. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat
1102 Methods. 2012;9(4): 357–9.
1103 34. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al.
1104 Full-length transcriptome assembly from RNA-Seq data without a reference
1105 genome. Nat Biotechnol. 2011;29(7): 644–52.
1106 35. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using
1107 DIAMOND. Nat Methods. 2015;12(1): 59–60.
1108 36. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment
1109 search tool. J Mol Biol. 1990;215(3): 403–10.
1110 37. Fauquet CM, Stanley J. Revising the way we conceive and name viruses below
1111 the species level: A review of geminivirus taxonomy calls for new standardized
1112 isolate descriptors. Arch Virol. 2005;150: 2151–2179.
1113 38. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al.
1114 Primer3-new capabilities and interfaces. Nucleic Acids Res. 2012;40(15):
1115 e115.
1116 39. Cameron SA, Hines HM, Williams PH. A comprehensive phylogeny of the
1117 bumble bees (*Bombus*). Biol J Linn Soc. 2007;91(1): 161–88.
1118 40. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence
1119 alignment, interactive sequence choice and visualization. Brief Bioinform.
1120 2017. doi: 10.1093/bib/bbx108

1121 41. Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: improvement in
1122 accuracy of multiple sequence alignment. Nucleic Acids Res. 2005;33(2): 511–
1123 518.
1124 42. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, et al. BEAST 2:
1125 A software platform for Bayesian evolutionary analysis. PLoS Comput Biol.
1126 2014;10(4): e1003537.
1127 43. Bouckaert RR, Drummond AJ. bModelTest: Bayesian phylogenetic site model
1128 averaging and model comparison. BMC Evol Biol. 2017;17(1): 42.
1129 44. Heled J, Drummond AJ. Calibrated tree priors for relaxed phylogenetics and
1130 divergence time estimation. Syst Biol. 2012;61(1): 138–49.
1131 45. Drummond AJ, Ho SY, Phillips MJ, Rambaut A. Relaxed phylogenetics and
1132 dating with confidence. PLoS Biol. 2006;4(5): 699–710.
1133 46. Hines HM. Historical biogeography, divergence times, and diversification
1134 patterns of bumble bees (Hymenoptera: Apidae: *Bombus*). Syst Biol.
1135 2008;57(1): 58–75.
1136 47. Rambaut A, Baele G, Xie D, Drummond AJ, Suchard MA. Tracer v1.7. 2017.
1137 Available from: http://beast.bio.ed.ac.uk/Tracer
1138 48. Kanehisa M, Goto S, Kawashima S, Nakaya A. The KEGG databases at
1139 GenomeNet. Nucleic Acids Res. 2002;30(1): 42–46.
1140 49. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al.
1141 Pfam: The protein families database. Nucleic Acids Res. 2014;42(D1): D222–
1142 D230.
1143 50. Kamer G, Argos P. Primary structural comparison of RNA-dependent
1144 polymerases from plant, animal and bacterial viruses. Nucleic Acids Res.
1145 1984;12(18): 7269–82.

1146 51. Pei J, Kim BH, Grishin NV. PROMALS3D: A tool for multiple protein
1147 sequence and structure alignments. Nucleic Acids Res. 2008;36(7): 2295–300.
1148 52. Koonin EV, Dolja VV, Krupovic M. Origins and evolution of viruses of
1149 eukaryotes: The ultimate modularity. Virology. 2015;79-80: 2-25.
1150 53. Zanotto PM, Gibbs MJ, Gould EA, Holmes EC. A reevaluation of the higher
1151 taxonomy of viruses based on RNA polymerases. J Virol. 1996;70(9): 6083–
1152 6096.
1153 54. Gorbalenya AE, Pringle FM, Zeddam JL, Luke BT, Cameron CE, Kalmakoff J,
1154 et al. The palm subdomain-based active site is internally permuted in viral
1155 RNA-dependent RNA polymerases of an ancient lineage. J Mol Biol.
1156 2002;324(1): 47–62.
1157 55. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein
1158 blocks. Proc Natl Acad Sci U S A. 1992;89(22): 10915–10919.
1159 56. Ferreira MAR, Suchard MA. Bayesian analysis of elapsed times in continuous-
1160 time Markov chains. Can J Stat. 2008;36(3): 355–68.
1161 57. Holmes EC. Molecular clocks and the puzzle of RNA virus origins. J Virol.
1162 2003;77(7): 3893–3897.
1163 58. Katzourakis A, Gifford RJ. Endogenous viral elements in animal genomes.
1164 PLoS Genet. 2010;6(11): e1001191.
1165 59. Chiba S, Kondo H, Tani A, Saisho D, Sakamoto W, Kanematsu S, et al.
1166 Widespread endogenization of genome sequences of non-retroviral RNA
1167 viruses into plant genomes. PLoS Pathog. 2011;7(7): e1002146.
1168 60. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with
1169 BEAUti and the BEAST 1.7. Mol Biol Evol. 2012;29(8): 1969–1973.
1170 61. Webster CL, Waldron FM, Robertson S, Crowson D, Ferrari G, Quintana JF, et

1171 al. The discovery, distribution, and evolution of viruses associated with
1172 *Drosophila melanogaster*. PLoS Biol. 2015;13(7): e1002210.
1173 62. Ebert TA, Brlansky R, Rogers M. Reexamining the pooled sampling approach
1174 for estimating prevalence of infected insect vectors. Ann Entomol Soc Am.
1175 2010;103(6): 827–37.
1176 63. Gibbs AJ, Gower JC. The use of a multiple-transfer method in plant virus
1177 transmission studies - Some statistical points arising in the analysis of results.
1178 Ann Appl Biol. 1960;48(1): 75–83.
1179 64. Thompson KH. Estimation of the proportion of vectors in a natural population
1180 of insects. Biometrics. 1962;18(4): 568–78.
1181 65. Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et
1182 al. Stan: A probabilistic programming language. J Stat Softw. 2017;76(1): 1–
1183 32.
1184 66. Liu Y, Gelman A, Zheng T. Simulation-efficient shortest probability intervals.
1185 Stat Comput. 2015;25(4): 809–819.
1186 67. Stan Development Team. Stan Modeling Language: User's Guide and
1187 Reference Manual. 2.17.1. 2017. Available from: https://github.com/stan-
1188 dev/stan/releases/download/v2.17.1/stan-reference-2.17.1.pdf.
1189 68. Goulson D, Hanley ME, Darvill B, Ellis JS, Knight ME. Causes of rarity in
1190 bumblebees. Biol Conserv. 2005;122(1): 1–8.
1191 69. Bailey L, Ball BV. Honey bee pathology. 2nd ed. London: Academic Press
1192 Inc.; 1991.
1193 70. de Miranda JR, Cornman RS, Evans JD, Semberg E, Haddad N, Neumann P, et
1194 al. Genome characterization, prevalence and distribution of a macula-like virus
1195 from *Apis mellifera* and *Varroa destructor*. Viruses. 2015;7(7): 3586–602.

1196 71. Schoonvaere K, Smet LD, Smagghe G, Vierstraete A, Braeckman BP, de Graaf
1197 DC, et al. Unbiased RNA shotgun metagenomics in social and solitary wild
1198 bees detects associations with eukaryote parasites and new viruses. PLoS One.
1199 2016;11(12): e0168456.
1200 72. Bailey L, Woods RD. Two more small RNA viruses from honey bees and
1201 further observations on sacbrood and acute bee-paralysis viruses. J Gen Virol.
1202 1977;37(1): 175–82.
1203 73. Remnant EJ, Shi M, Buchmann G, Blacquière T, Holmes EC, Beekman M, et
1204 al. A diverse range of novel RNA viruses in geographically distinct honey bee
1205 populations. J Virol. 2017; 91(16): e00158-17.
1206 74. Runckel C, Flenniken ML, Engel JC, Ruby JG, Ganem D, Andino R, et al.
1207 Temporal analysis of the honey bee microbiome reveals four novel viruses and
1208 seasonal prevalence of known viruses, *Nosema*, and *Crithidia*. PLoS One.
1209 2011;6(6): e20656.
1210 75. Boccardo G, Milne RG, Luisoni E, Lisa V, Accotto GP. Three seedborne
1211 cryptic viruses containing double-stranded RNA isolated from white clover.
1212 Virology. 1985;147(1): 29–40.
1213 76. Smith KM, Markham R. Two new viruses affecting tobacco and other plants.
1214 Phytopathology. 1944;34: 324–329.
1215 77. Koloniuk I, Přibylová J, Fránová J. Molecular characterization and complete
1216 genome of a novel nepovirus from red clover. Arch Virol. 2018;163(5): 1387–
1217 9.
1218 78. Bronkhorst AW, van Cleef KW, Vodovar N, İnce İA, Blanc H, Vlak JM, et al.
1219 The DNA virus Invertebrate iridescent virus 6 is a target of the *Drosophila*
1220 RNAi machinery. Proc Natl Acad Sci U S A. 2012;109(51): E3604–13.

1221 79. Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, et al.
1222 Discrete small RNA-generating loci as master regulators of transposon activity
1223 in *Drosophila*. Cell. 2007;128(6): 1089–103.
1224 80. Suzuki Y, Frangeul L, Dickson LB, Blanc H, Verdier Y, Vinh J, et al.
1225 Uncovering the repertoire of endogenous flaviviral elements in Aedes. J Virol.
1226 2017;91(15): e00571-17.
1227 81. Lewis SH, Quarles KA, Yang Y, Tanguy M, Frézal L, Smith SA, et al. Pan-
1228 arthropod analysis reveals somatic piRNAs as an ancestral defence against
1229 transposable elements. Nat Ecol Evol. 2018;2(1): 174–81.
1230 82. van Rij RP, Saleh MC, Berry B, Foo C, Houk A, Antoniewski C, et al. The
1231 RNA silencing endonuclease Argonaute 2 mediates specific antiviral immunity
1232 in *Drosophila melanogaster*. Genes Dev. 2006;20(21): 2985–95.
1233 83. van Mierlo JT, Bronkhorst AW, Overheul GJ, Sadanandan SA, Ekström JO,
1234 Heestermans M, et al. Convergent evolution of Argonaute-2 Slicer antagonism
1235 in two distinct insect RNA viruses. PLoS Pathog. 2012;8(8): e1002872.
1236 84. Deleris A, Gallago-Bartolome J, Bao J, Kasschau KD, Carrington JC, Voinnet
1237 O. Hierarchical action and inhibition of plant dicer-like proteins in antiviral
1238 defense. Science. 2006;313(5783): 68–71.
1239 85. Mordecai GJ, Wilfert L, Martin SJ, Jones IM, Schroeder DC. Diversity in a
1240 honey bee pathogen: first report of a third master variant of the Deformed
1241 Wing Virus quasispecies. ISME J. 2015;10(5): 1–10.
1242 86. Schoonvaere K, Smagghe G, Francis F, de Graaf DC. Study of the
1243 metatranscriptome of eight social and solitary wild bee species reveals novel
1244 viruses and bee parasites. Front Microbiol. 2018. doi:
1245 10.3389/fmicb.2018.00177

1246 87. Roberts JMK, Anderson DL, Durr PA. Metagenomic analysis of Varroa-free
1247 Australian honey bees (*Apis mellifera*) shows a diverse Picornavirales virome.
1248 J Gen Virol. 2018;99: 818–26.
1249 88. Félix MA, Ashe A, Piffaretti J, Wu G, Nuez I, Bélicard T, et al. Natural and
1250 experimental infection of *Caenorhabditis* nematodes by novel viruses related to
1251 nodaviruses. PLoS Biol. 2011;9(1): e1000586.
1252 89. Rao S, Poinar G, Henley D. A scientific note on rare parasitism of the bumble
1253 bee pollinator, *Bombus impatiens*, by a mermithid nematode, *Pheromermis* sp.
1254 (Nematoda: Mermithidae). Apidologie. 2017;48(1): 75–7.
1255 90. Wang XB, Jovel J, Udomporn P, Wang Y, Wu Q, Li WX, et al. The 21-
1256 nucleotide, but not 22-nucleotide, viral secondary small interfering RNAs
1257 direct potent antiviral defense by two cooperative Argonautes in *Arabidopsis*
1258 *thaliana*. Plant Cell. 2011;23(4): 1625–38.
1259 91. Longdon B, Hadfield JD, Day JP, Smith SC, McGonigle JE, Cogni R, et al.
1260 The causes and consequences of changes in virulence following pathogen host
1261 shifts. PLoS Pathog. 2015;11(3): e1004728.
1262 92. Roberts K, Hadfield JD, Sharma MD, Longdon B. Changes in temperature alter
1263 susceptibility to a virus following a host shift. PLoS Patho. 2018;14(10):
1264 e1007185
1265 93. Waxman D, Weinert LA, Welch JJ. Inferring host range dynamics from
1266 comparative data: The protozoan parasites of New World monkeys. Am Nat.
1267 2014;184(1): 65–74.
1268 94. Bandín I, Dopazo CP. Host range, host specificity and hypothesized host shift
1269 events among viruses of lower vertebrates. Vet Res. 2011;42(67).
1270 95. Langsjoen RM, Haller SL, Roy CJ, Vinet-Oliphant H, Bergren NA, Erasmus

1271 JH, et al. Chikungunya virus strains show lineage-specific variations in
1272 virulence and cross-protective ability in murine and nonhuman primate models.
1273 MBio. 2018;9(2): e02449-17.
1274 96. Baker C, Antonovics J. Evolutionary determinants of genetic variation in
1275 susceptibility to infectious diseases in humans. PLoS One. 2012; 7(1): e29089.
1276 97. Davies TJ, Pedersen AB. Phylogeny and geography predict pathogen
1277 community similarity in wild primates and humans. Proc R Soc B Biol Sci.
1278 2008;275(1643): 1695–701.
1279 98. Huang S, Bininda-Emonds OR, Stephens PR, Gittleman JL, Altizer S.
1280 Phylogenetically related and ecologically similar carnivores harbour similar
1281 parasite assemblages. J Anim Ecol. 2014;83(3): 671–80.
1282 99. Cooper N, Griffin R, Franz M, Omotayo M, Nunn CL, Fryxell J. Phylogenetic
1283 host specificity and understanding parasite sharing in primates. Ecol Lett.
1284 2012;15(12): 1370–7.
1285 100. Longdon B, Day JP, Alves JM, Smith SCL, Houslay TM, McGonigle JE, et al.
1286 Host shifts result in parallel genetic changes when viruses evolve in closely
1287 related species. PLoS Pathog. 2018;14(4): e1006951.
1288 101. Streicker DG, Altizer SM, Velasco-Villa A, Rupprecht CE. Variable
1289 evolutionary routes to host establishment across repeated rabies virus host
1290 shifts among bats. Proc Natl Acad Sci U S A. 2012;109(48): 19715-19720.
1291 102. Durrer S, Schmid-Hempel P. Shared use of flowers leads to horizontal
1292 pathogen transmission. Proc R Soc B Biol Sci. 1994;258(1353): 299–302.
1293 103. Graystock P, Goulson D, Hughes WO. Parasites in bloom: flowers aid
1294 dispersal and transmission of pollinator parasites within and between bee
1295 species. Proc R Soc B Biol Sci. 2015;282(1813): 20151371.

1296 104. McArt SH, Koch H, Irwin RE, Adler LS. Arranging the bouquet of disease:
1297 floral traits and the transmission of plant and animal pathogens. Ecology
1298 Letters. 2014;17(5): 624-36.
1299 105. Ruiz-González MX, Bryden J, Moret Y, Reber-Funk C, Schmid-Hempel P,
1300 Brown MJ. Dynamic transmission, host quality, and population structure in a
1301 multihost parasite of bumblebees. Evolution. 2012;66(10): 3053–66.
1302 106. Inouye DW. Resource partitioning in bumblebees: Experimental studies of
1303 foraging behavior. Ecology. 1978;59(4): 672–8.
1304 107. Koonin EV, Wolf YI, Nagasaki K, Dolja VV. The Big Bang of picorna-like
1305 virus evolution antedates the radiation of eukaryotic supergroups. Nat Rev
1306 Microbiol. 2008;6(12): 925–39.
1307 108. Fürst MA, McMahon DP, Osborne JL, Paxton RJ, Brown MJ. Disease
1308 associations between honeybees and bumblebees as a threat to wild pollinators.
1309 Nature. 2014;506(7488): 364–366.
1310 109. McMahon DP, Fürst MA, Caspar J, Theodorou P, Brown MJF, Paxton RJ. A
1311 sting in the spit: Widespread cross-infection of multiple RNA viruses across
1312 wild and managed bees. J Anim Ecol. 2015;84(3): 615–24.
1313