The evolution, diversity and host associations of rhabdoviruses

Ben Longdon[1*], Gemma GR Murray[1], William J Palmer[1], Jonathan P Day[1], Darren J Parker[2, 3], John J Welch[1], Darren J Obbard[4] and Francis M Jiggins[1].

[1]Department of Genetics
University of Cambridge
Cambridge
CB2 3EH
UK

[2] School of Biology
University of St. Andrews
St. Andrews
KY19 9ST
UK

[3]Department of Biological and Environmental Science,
University of Jyväskylä,
Jyväskylä,
Finland

[4]Institute of Evolutionary Biology, and Centre for Immunity Infection and Evolution
University of Edinburgh
Edinburgh
EH9 3JT
UK

*corresponding author
email: b.longdon@gen.cam.ac.uk
phone: +441223333945

Abstract

Metagenomic studies are leading to the discovery of a hidden diversity of RNA viruses, but new approaches are needed predict the host species these poorly characterised viruses pose a risk to. The rhabdoviruses are a diverse family of RNA viruses that includes important pathogens of humans, animals and plants. We have discovered the sequences of 32 new rhabdoviruses through a combination of our own RNA sequencing of insects and searching public sequence databases. Combining these with previously known sequences we reconstructed the phylogeny of 195 rhabdovirus sequences producing the most in depth analysis of the family to date. In most cases we know nothing about the biology of the viruses beyond the host they were identified from, but our dataset provides a powerful phylogenetic approach predict which are vector-borne viruses and which are specific to vertebrates or arthropods. By reconstructing ancestral and present host states we found that switches between major groups of hosts have occurred rarely during rhabdovirus evolution. This allowed us to propose that 76 new likely vector-borne vertebrate viruses among viruses identified from vertebrates or biting insects. Our analysis suggests it is likely there was a single origin of the plant viruses and arthropod-borne vertebrate viruses while vertebrate-specific viruses arose at least twice. There are also two large clades of viruses that infect insects, including the sigma viruses, which are vertically transmitted. There are also few transitions between aquatic and terrestrial ecosystems. Viruses also cluster together at a finer scale, with closely related viruses tending to be found in closely related hosts. Our data therefore suggest that throughout their evolution rhabdoviruses have occasionally made a long distance host jump before spreading through related hosts in the same environment. This approach offers a way to predict the most probable biology and key traits of newly discovered viruses.

Keywords
Virus, Host shift, Arthropod, Insect, Rhabdoviridae, Mononegavirales

Introduction

RNA viruses are an abundant and diverse group of pathogens. In the past, viruses were typically isolated from hosts displaying symptoms of infection, before being characterized morphologically and then sequenced following PCR [1, 2]. PCR-based detection of novel RNA viruses is problematic as there is no single conserved region of the genome of viruses from a single family, let alone all RNA viruses. High throughput next generation sequencing technology has revolutionized virus discovery, allowing rapid detection and sequencing of divergent virus sequences simply by sequencing total RNA from infected individuals [1, 2]

One particularly diverse family of RNA viruses is the *Rhabdoviridae*. Rhabdoviruses are negative-sense single-stranded RNA viruses in the order *Mononegavirales* [3]. They infect an extremely broad range of hosts and have been discovered in plants, fish, mammals, reptiles and a broad range of insects and other arthropods [4]. The family includes important pathogens of humans and livestock. Perhaps the most well-known is rabies virus, which can infect a diverse array of mammals and causes a fatal infection killing 59,000 people per year with an estimated economic cost of $8.6 billion (US) [5]. Other rhabdoviruses such as vesicular stomatitis virus and bovine ephemeral fever virus are important pathogens of domesticated animals, whilst others are pathogens of crops [3].

Arthropods play a key role in transmission of many rhabdoviruses. Many viruses found in vertebrates have also been detected in arthropods, including sandflies, mosquitoes, ticks and midges [6]. The rhabdoviruses that infect plants are also often transmitted by arthropods [7] and the rhabdoviruses that infect fish have the potential to be vectored by ectoparasitic copepod sea-lice [8, 9]. Rhabdoviruses replicate upon infection of insects vectors (insects are not just mechanical vectors) [7]. Other rhabdoviruses are insect-specific. In particular, the sigma viruses are a clade of vertically transmitted viruses that infect dipterans and are well-studied in *Drosophila* [10-12]. Recently, a number of rhabdoviruses have been found to be associated with a wide array of insect and other arthropod species, suggesting they may be common arthropod viruses [13, 14]. Furthermore, a number of arthropod genomes contain integrated endogenous viral elements (EVEs) with similarity to rhabdoviruses, suggesting that these species have been infected with rhabdoviruses [15-18].

Here we aimed to uncover the diversity of the rhabdoviruses, and examine how they have switched between different host taxa during their evolutionary history. Insects infected with rhabdoviruses commonly become paralysed on exposure to CO_2 [19-21]. We exploited this fact to screen field collections of flies from several continents for novel rhabdoviruses that were then sequenced using RNA-sequencing (RNA-seq). Additionally we searched for rhabdovirus-like sequences in publicly available RNA-seq data. We identified 32 novel rhabdovirus-like sequences from a wide array of invertebrates and plants, and combined them with recently discovered viruses to produce the most comprehensive phylogeny of the rhabdoviruses to date. For many of the viruses we do not know their true host range, so we used the phylogeny to identify a large number of

110 new likely vector-borne viruses and to reconstruct the evolutionary history of this
111 diverse group of viruses.
112
113
114 **Methods**
115
116 *Discovery of new rhabdoviruses by RNA sequencing*
117
118 Diptera (flies, mostly Drosophilidae) were collected in the field from Spain, USA, Kenya,
119 France, Ghana and the UK (Data S1: http://dx.doi.org/10.6084/m9.figshare.1425432).
120 Infection with rhabdoviruses can cause *Drosophila* and other insects to become
121 paralysed after exposure to CO_2 [19-21], so we enriched our sample for infected
122 individuals by exposing them to CO_2 at 12°C for 15 mins, only retaining individuals that
123 showed symptoms of paralysis 30mins later. We extracted RNA from 79 individual
124 insects (details in Data S1 http://dx.doi.org/10.6084/m9.figshare.1425432) using
125 Trizol reagent (Invitrogen) and combined the extracts into two pools (retaining non-
126 pooled individual RNA samples). RNA was then rRNA depleted with the Ribo-Zero Gold
127 kit (epicenter, USA) and used to construct Truseq total RNA libraries (Illumina).
128 Libraries were constructed and sequenced by BGI (Hong Kong) on an Illumina Hi-Seq
129 2500 (one lane, 100bp paired end reads, generating ~175 million reads). Sequences
130 were quality trimmed with Trimmomatic (v3); Illumina adapters were clipped, bases
131 were removed from the beginning and end of reads if quality dropped below a
132 threshold, sequences were trimmed if the average quality within a window fell below a
133 threshold and reads less than 20 base pairs in length were removed. We *de novo*
134 assembled the RNA-seq reads with Trinity (release 2013-02-25) using default settings
135 and jaccard clip option for high gene density. The assembly was then blasted (tblastn)
136 with rhabdovirus coding sequences as the query to identify rhabdovirus-like sequences.
137 Contigs with hits were then reciprocally comapred against Genbank cDNA and RefSeq
138 nucleotide databases using tblastn and only retained if they hit a virus-like sequence.
139 Raw read data were deposited in the NCBI Sequence Read Archive (SRP057824).
140 Putative viral sequences have been submitted to Genbank (accessions in Tables S1 and
141 S2 http://dx.doi.org/10.6084/m9.figshare.1502665).
142
143 As the RNA-seq was performed on pooled samples, we assigned rhabdovirus sequences
144 to individual insects by PCR on individual samples RNA. cDNA was produced using
145 Promega GoScript Reverse Transcriptase and random-hexamer primers, and PCR
146 performed using primers designed using the rhabdovirus sequences. Infected host
147 species were identified by sequencing the mitochondrial gene *COI*. We were unable to
148 identify the host species of the virus from a *Drosophila affinis* sub-group species
149 (sequences appear similar to both *Drosophila affinis* and the closely related Drosophila
150 *athabasca*), despite the addition of further mitochondrial and nuclear sequences to
151 increase confidence. In all cases we confirmed that viruses were only present in cDNA
152 and not in non reverse-transcription (RT) controls (i.e. DNA) by PCR, and so they cannot
153 be integrated into the insect genome (i.e. endogenous virus elements or EVEs [17]). *COI*
154 primers were used as a positive control for the presence of DNA in the non RT template.
155

156 We identified sigma virus sequences in RNA-seq data from *Drosophila montana* [22]. We
157 used RT-PCR on an infected fly line to amplify the virus sequence, and carried out
158 additional Sanger sequencing with primers designed using the RNA-seq assembly.
159 Additional virus sequences were identified from an RNA-seq analysis of pools of wild
160 caught *Drosophila*: DImmSV from *Drosophila immigrans* (collection and sequencing
161 described [23]), DTriSV from a pool of *Drosophila tristis* and SDefSV from
162 *Scaptodrosophila deflexa* (both Darren Obbard, unpublished data), accessions in tables
163 S1 and S2 (http://dx.doi.org/10.6084/m9.figshare.1502665).
164
165 *Discovery of rhabdoviruses in public sequence databases*
166
167 Rhabdovirus L gene sequences were used to search (tblastn) against expressed
168 sequence tag (EST) and transcriptome shotgun assembly (TSA) databases (NCBI). All
169 hits were reciprocally blasted against Genbank cDNA and RefSeq databases and only
170 retained if they hit a virus-like sequence. We used two approaches to examine whether
171 sequences were present as RNA but not DNA. First, where assemblies of whole-genome
172 shotgun sequences were available, we used BLAST to test whether sequences were
173 integrated into the host genome. Second, for the virus sequences in the butterfly *Pararge*
174 *aegeria* and the medfly *Ceratitis capitata* we were able to obtain infected samples to
175 confirm the sequences are only present in RNA by performing PCR on both genomic
176 DNA and cDNA as described above (samples kindly provided by Casper
177 Breuker/Melanie Gibbs, and Philip Leftwich respectively)
178
179 *Phylogenetic analysis*
180
181 All available rhabdovirus-like sequences were downloaded from Genbank (accessions in
182 Data S2: http://dx.doi.org/10.6084/m9.figshare.1425419). Amino acid sequences for
183 the L gene (encoding the RNA Dependent RNA Polymerase or RDRP) were used to infer
184 the phylogeny (L gene sequences: http://dx.doi.org/10.6084/m9.figshare.1425067), as
185 they contain conserved domains that can be aligned across this diverse group of viruses.
186 Sequences were aligned with MAFFT [24] under default settings and then poorly aligned
187 and divergent sites were removed with either TrimAl (v1.3 strict settings, implemented
188 on Phylemon v2.0 server, alignment: http://dx.doi.org/10.6084/m9.figshare.1425069)
189 [25] or Gblocks (v0.91b selecting smaller final blocks, allowing gap positions and less
190 strict flanking positions to produce a less stringent selection, alignment:
191 http://dx.doi.org/10.6084/m9.figshare.1425068) [26]. These resulted in alignments of
192 1492 And 829 amino acids respectively.
193
194 Phylogenetic trees were inferred using Maximum Likelihood in PhyML (v3.0) [27] using
195 the LG substitution model [28] (preliminary analysis confirmed the results were robust
196 to the amino acid substitution model selected), with a gamma distribution of rate
197 variation with four categories and using a sub-tree pruning and regrafting topology
198 searching algorithm. Branch support was estimated using Approximate Likelihood-Ratio
199 Tests (aLRT) that is reported to outperform bootstrap methods [29]. Figures were
200 created using FIGTREE (v. 1.4) [30].
201
202 *Analysis of genetic structure between viruses taken from different hosts and ecologies*

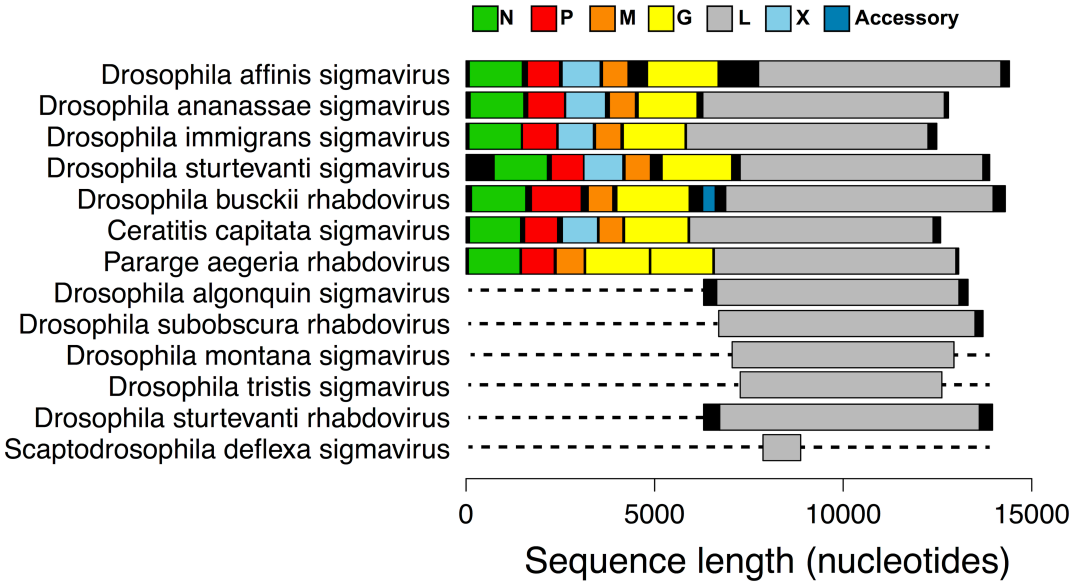
203
204 We measured the degree of genetic structure between virus sequences identified in
205 different categories of host (arthropods, vertebrates and plants) and ecosystems
206 (terrestrial and aquatic). We measured the degree of genetic structure between virus
207 sequences from different groups of hosts/ecosystems using Hudson's F_{st} estimator [31]
208 following the recommendations of [32]. We calculated F_{st} as:

209 $$F_{st} = 1 - \left(\frac{\text{number of differences within a population}}{\text{number of differences between populations}} \right)$$

210 where a population is a host category or ecosystem. The significance of this value was
211 tested by permuting the host/ecosystem association of viruses. Since arthropods are the
212 most sampled host, we tested for evidence of genetic structure within the arthropod-
213 associated viruses that would suggest co-divergence with their hosts or preferential
214 host-switching between closely related hosts. We calculated the Pearson correlation
215 coefficient of the evolutionary distances between viruses and the evolutionary distances
216 between their hosts and tested for significance by permutation (as in [33]). We used the
217 patristic distances of our ML tree for the virus data and a time-tree of arthropod genera,
218 using published estimates of divergence dates [34, 35].
219
220 *Reconstruction of host associations*
221
222 Viruses were categorised as having one of four types of host association: arthropod-
223 specific, vertebrate-specific, arthropod-vectored plant, or arthropod-vectored
224 vertebrate. However, the host association of some viruses are uncertain when they have
225 been isolated from vertebrates, biting-arthropods or plant-sap-feeding arthropods. Due
226 to limited sampling it was not clear whether viruses isolated from vertebrates were
227 vertebrate specific or arthropod-vectored vertebrate viruses; or whether viruses
228 isolated from biting-arthropods were arthropod specific viruses or arthropod-vectored
229 vertebrate viruses; or if viruses isolated from plant-sap-feeding arthropods were
230 arthropod-specific or arthropod-vectored plant viruses.
231
232 We omitted three viruses that were isolated from hosts outside of these four categories
233 (viruses from a nematode, fungus and cnidarian) from our analyses. We classified three
234 of the fish infecting dimarhabdoviruses as vertebrate specific based on the fact they can
235 be transmitted via immersion in water containing virus during experimental conditions
236 [36-38], and the widely held belief amongst the fisheries community that these viruses
237 are not typically vectored [8]. However, there is some evidence these viruses can be
238 transmitted by arthropods (sea lice) in experiments [8, 9] and so we would recommend
239 this be interpreted with some caution. Additionally, although we classified the viruses
240 identified in sea-lice as having biting arthropod hosts, they may be crustacean-specific.
241 The two viruses from *Lepeophtheirus salmonis* do not seem to infect the fish they
242 parasitise and are present in all developmental stages of the lice, suggesting they may be
243 transmitted vertically [39].
244
245 We simultaneously estimated both the current and ancestral host associations, and the
246 phylogeny of the viruses, using a Bayesian analysis, implemented in BEAST v1.8 [40, 41].
247 Since accurate branch lengths are essential for this analysis, we used a subset of the

248 sites and strains used in the Maximum Likelihood (ML) analysis. We retained 189 taxa;
249 all rhabdoviruses excluding the divergent fish-infecting novirhabdovirus clade and the
250 virus from *Hydra*, as well as the viruses from *Lolium perenne and Conwentzia*
251 *psociformis*, which had a large number of missing sites. Sequences were trimmed to a
252 conserved region of 414 amino acids where data was recorded for most of these viruses
253 (the Gblocks alignment trimmed further by eye:
254 http://dx.doi.org/10.6084/m9.figshare.1425431). We used the host-association
255 categories described above, which included ambiguous states. To model amino acid
256 evolution we used an LG substitution model with gamma distributed rate variation
257 across sites [28] and an uncorrelated lognormal relaxed clock model of rate variation
258 among lineages [42]. To model the evolution of the host associations we used an
259 asymmetric transition rate matrix (allowing transitions to and from a host association to
260 take place at different rates) and a strict clock model. We also examined how often these
261 viruses jumped between different classes of hosts using reconstructed counts of
262 biologically feasible changes of host association and their HPD confidence intervals (CI)
263 using Markov Jumps [43]. These included switches between arthropod-specific and both
264 arthropod-vectored vertebrate and arthropod-vectored plant states, and between
265 vertebrate specific and arthropod-vectored vertebrate states. We used a constant
266 population size coalescent prior for the relative node ages (using a birth-death prior
267 gave equivalent results) and the BEAUti v1.8 default priors for all other parameters [40]
268 (BEAUti xml http://dx.doi.org/10.6084/m9.figshare.1431922). Convergence was
269 assessed using Tracer v1.6 [44], and a burn-in of 30% was removed prior to the
270 construction of a consensus tree, which included a description of ancestral host
271 associations. High effective sample sizes were achieved for all parameters (>200).
272
273 The maximum clade credibility tree estimated
274 (http://dx.doi.org/10.6084/m9.figshare.1425436) for the host association
275 reconstruction was very similar to the independently estimated maximum likelihood
276 phylogeny (http://dx.doi.org/10.6084/m9.figshare.1425083), which made no
277 assumptions about the appropriateness or otherwise of applying a clock model. The
278 minor topological differences may be expected in reconstructions that differ in their
279 assumptions about evolutionary rates. In Figure 2 we have transferred the ancestral
280 state reconstruction from the BEAST tree to the maximum likelihood tree.
281
282
283 **Results**
284
285 *Novel rhabdoviruses from RNA-seq*
286
287 To search for new rhabdoviruses we collected a variety of different species of flies,
288 screened them for CO_2 sensitivity, which is a common symptom of infection, and
289 sequenced total RNA of these flies by RNA-seq. We identified rhabdovirus-like
290 sequences from a *de-novo* assembly by BLAST, and used PCR to identify which samples
291 these sequences came from.
292
293 This approach resulted in eleven rhabdovirus-like sequences from nine (possibly ten)
294 species of fly. Seven of these viruses were previously unknown and four had been

295 reported previously from shorter sequences (Tables S1 and S2
296 http://dx.doi.org/10.6084/m9.figshare.1502665). The novel viruses were highly
297 divergent from known viruses. Sigma viruses known from other species of *Drosophila*
298 typically have genomes of ~12.5Kb [12, 45], and six of our sequences were
299 approximately this size, suggesting they are near-complete genomes. None of the
300 viruses discovered in our RNA-seq data were integrated into the host genome (see
301 Methods for details).
302
303 To investigate the putative gene content of the viruses, we predicted putative genes
304 based on open reading frames (ORFs). For the viruses with apparently complete
305 genomes (Figure 1), we found that those from *Drosophila ananassae, Drosophila affinis,*
306 *Drosophila immigrans* and *Drosophila sturtvanti* contained ORFs corresponding to the
307 five core genes found across all rhabdoviruses, with an additional ORF between the P
308 and M genes. This is the location of the X gene found in sigma viruses, and in three of the
309 four novel viruses it showed BLAST sequence similarity to the X gene of sigma viruses.
310 The virus from *Drosophila busckii* did not contain an additional ORF between the P and
311 M genes, but instead contained an ORF between the G and L gene. The high mutation
312 rate of RNA viruses mean that such ORFs are highly likely to be functional and retained
313 by selection, as non-function ORFs would quickly be degraded.
314
315 Using the phylogeny described below, we have classified our newly discovered viruses
316 as either sigma viruses, rhabdoviruses or other viruses, and named them after the host
317 species they were identified from (Figure 1) [46]. We also found one other novel
318 mononegavirales-like sequence from *Drosophila unispina* that groups with a recently
319 discovered clade of arthropod associated viruses (Nyamivirus clade [13], see Table S5
320 and the full phylogeny: http://dx.doi.org/10.6084/m9.figshare.1425083), as well as five
321 other RNA viruses from various families (data not shown), confirming our approach can
322 detect a wide range of divergent viruses.
323



324
325 **Figure 1. Genome organization of newly discovered viruses.**
326 Putative genes are shown in colour, non-coding regions are shown in black. ORFs were
327 designated as the first start codon following the transcription termination sequence (7 U's) of the

328 previous ORF to the first stop codon. Dotted lines represent parts of the genome not sequenced.
329 These viruses were either from our own RNA-seq data, or were first found in in public databases
330 and key features verified by PCR and Sanger sequencing. Rhabdovirus genomes are typically
331 ~11-13kb long and contain five core genes 3'-N-P-M-G-L-5' [3]. However, a number of groups of
332 rhabdoviruses contain additional accessory genes and can be up to ~16kb long [14, 47].
333
334 *New rhabdoviruses from public databases*
335
336 We identified a further 26 novel rhabdovirus-like sequences by searching public
337 databases of assembled RNA-seq data with BLAST. These included 19 viruses from
338 arthropods (Fleas, Crustacea, Lepidoptera, Diptera), one from a Cnidarian (*Hydra*) and 5
339 from plants (Table S3). Of these viruses, 19 had sufficient amounts of coding sequence
340 (>1000bp) to include in the phylogenetic analysis (Table S3), whilst the remainder were
341 too short (Table S4).
342
343 Four viruses from databases had near-complete genomes based on their size. These
344 were from the moth *Triodia sylvina*, the house fly *Musca domestica* (99% nucleotide
345 identity to Wuhan house fly virus 2 [13]), the butterfly *Pararge aegeria* and the medfly
346 *Ceratitis capitata*, all of which contain ORFs corresponding to the five core rhabdovirus
347 genes. The sequence from *C. capitata* had an additional ORF between the P and M genes
348 with BLAST sequence similarity to the X gene in sigma viruses. There were several
349 unusual sequences. Firstly, in the virus from *P. aegeria* there appear to be two full-
350 length glycoprotein ORFs between the M and L genes (we confirmed by Sanger
351 sequencing that both exist and the stop codon between the two genes was not an error).
352 Secondly, the *Agave tequilana* transcriptome contained a L gene ORF on a contig that
353 was the length of a typical rhabdovirus genome but did not appear to contain typical
354 gene content, suggesting it has very atypical genome organization, or has been
355 misassembled, or is integrated into its host plant genome [48]. Finally, the virus from
356 *Hydra magnipapillata* contained six predicted genes, but the L gene (RDRP) ORF was
357 unusually long. Some of the viruses we detected may be EVEs inserted into the host
358 genome and subsequently expressed [18]. For example, this is likely the case for the
359 sequence from the silkworm *Bombyx mori* that we also found in the silkworm genome,
360 and the L gene sequence from *Spodoptera exigua* that contains stop codons. Under the
361 assumption that viruses integrated into host genomes once infected those hosts, this
362 does not affect our conclusions below about the host range of these viruses [15-17]. We
363 also found nine other novel mononegavirale-like sequences that group with recently
364 discovered clades of insect viruses [13] (see Table S5 and
365 http://dx.doi.org/10.6084/m9.figshare.1425083).
366
367 *Rhabdovirus Phylogeny*
368
369 To reconstruct the evolution of the *Rhabdoviridae* we have produced the most complete
370 phylogeny of the group to date (Figure 2). We aligned the relatively conserved L gene
371 (RNA Dependant RNA Polymerase) from our newly discovered viruses with sequences
372 of known rhabdoviruses to give an alignment of 195 rhabdoviruses (and 26 other
373 mononegavirales as an outgroup). We reconstructed the phylogeny using different
374 sequence alignments and methodologies, and these all gave qualitatively similar results

375 with the same major clades being reconstructed (Gblocks:
376 http://dx.doi.org/10.6084/m9.figshare.1425083, TrimAl:
377 http://dx.doi.org/10.6084/m9.figshare.1425082 and BEAST:
378 http://dx.doi.org/10.6084/m9.figshare.1425436). The ML and Bayesian relaxed clock
379 phylogenies were very similar, suggesting our analysis is robust to the assumptions of a
380 relaxed molecular clock. The branching order between the clades in the
381 dimarhabdovirus supergroup was generally poorly supported and differed between the
382 methods and alignments. Eight sequences that we discovered were not included in this
383 analysis as they were considered too short, but their closest BLAST hits are listed in
384 Table S4.
385
386 We recovered all of the major clades described previously (Figure 2), and found that the
387 majority of known rhabdoviruses belong to the dimarhabdovirus clade (Figure 2b). The
388 RNA-seq viruses from *Drosophila* fall into either the sigma virus clade (Figure 2b) or the
389 arthropod clade sister to the cyto- and nucleo- rhabdoviruses (Figure 2a). The viruses
390 from sequence databases are diverse, coming from almost all of the major clades with
391 the exception of the lyssaviruses.
392
393 *Predicted host associations of viruses*
394
395 With a few exceptions, rhabdoviruses are either arthropod-vectored viruses of plants or
396 vertebrates, or are vertebrate- or arthropod- specific. In many cases the only
397 information about a virus is the host from which it was isolated. Therefore, *a priori*, it is
398 not clear whether viruses isolated from vertebrates are vertebrate-specific or
399 arthropod-vectored, or whether viruses isolated from biting arthropods (e.g.
400 mosquitoes, sandflies, ticks, midges and sea lice) are arthropod specific or also infect
401 vertebrates. Likewise, it is not clear whether viruses isolated from sap-sucking insects
402 (all Hemiptera: aphids, leafhoppers, scale insect and mealybugs) are arthropod-specific
403 or arthropod-vectored plant viruses.
404
405 By combining data on the ambiguous and known host associations with phylogenetic
406 information, we are able to predict both the ancestral and present host associations of
407 viruses (http://dx.doi.org/10.6084/m9.figshare.1425436). To do this we used a
408 Bayesian phylogenetic analysis that simultaneously estimated the phylogeny and host
409 association of our data. In the analysis we defined our host associations either as
410 vertebrate-specific, arthropod-specific, arthropod-vectored vertebrate and arthropod-
411 vectored plant, or as ambiguous between two of these four states (see Methods).
412
413 This approach identified a large number of viruses that are likely to be new arthropod-
414 vectored vertebrate viruses (Figure 2b). 80 of 89 viruses with ambiguous host
415 associations were assigned a host association with strong posterior support (>0.95). Of
416 the 52 viruses found in biting arthropods, 45 were predicted to be arthropod-vectored
417 vertebrate viruses, and 6 to be arthropod-specific. Of the 30 viruses found in
418 vertebrates, 22 were predicted to be arthropod-vectored vertebrate viruses, and 2 were
419 predicted to be vertebrate-specific (both fish viruses). Of the 7 viruses found in plant-
420 sap-feeding arthropods (Figure 2a), 3 were predicted to be plant-associated and 2
421 arthropod-associated.

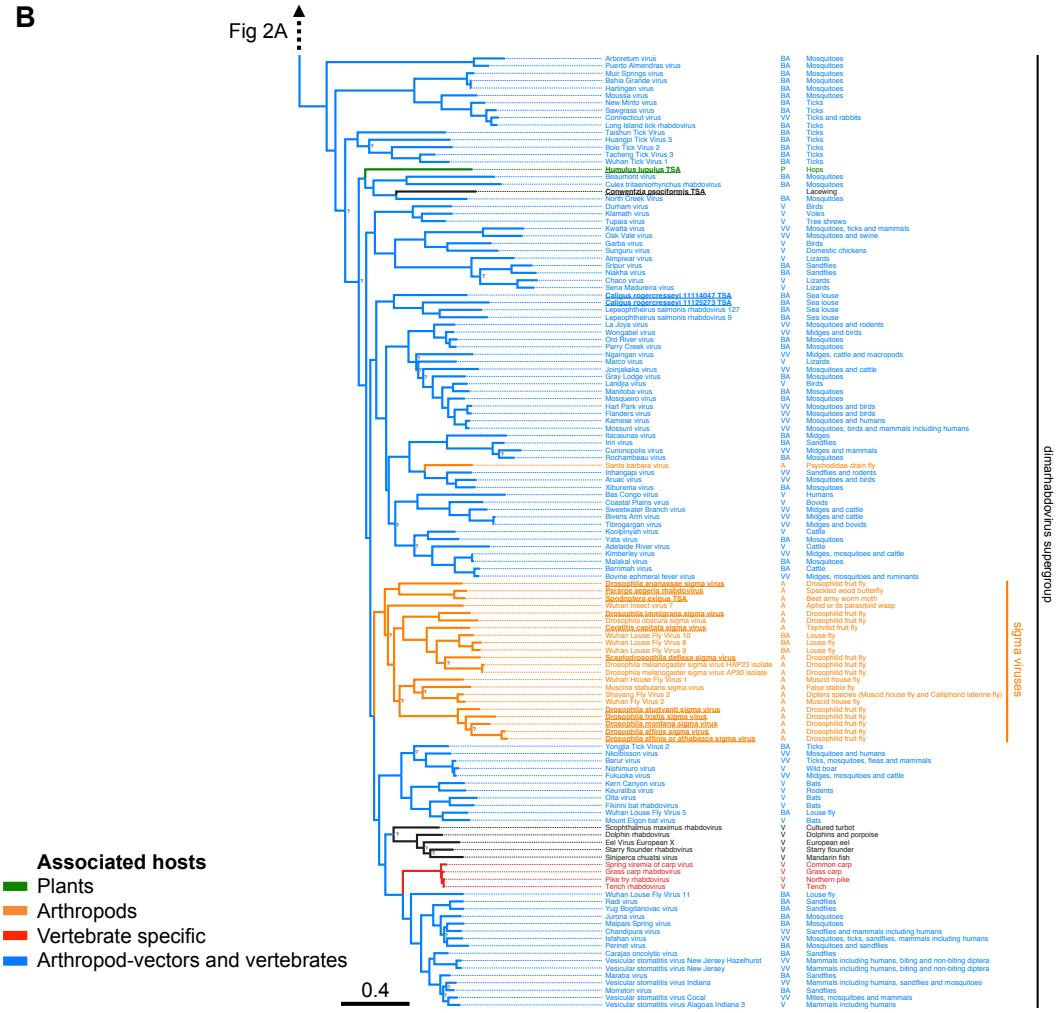


Figure 2. Maximum likelihood phylogeny of the *Rhabdoviridae*.

425 **(A)** shows the basal fish-infecting novirhabdoviruses, an unassigned group of arthropod
426 associated viruses, the plant infecting cyto- and nucleo- rhabdoviruses, as well as the vertebrate
427 specific lyssaviruses. **(B)** shows the dimarhabdovirus supergroup, which is predominantly
428 composed of arthropod-vectored vertebrate viruses, along with the arthropod-specific sigma
429 virus clade. Branches are coloured based on the Bayesian host association reconstruction
430 analysis. Black represents taxa omitted from host-state reconstruction or associations with <0.95
431 support. The tree was inferred from L gene sequences using the Gblocks alignment. The columns
432 of text are the virus name, the host category used for reconstructions, and known hosts (from left
433 to right). Codes for the host categories are: vs= vertebrate-specific, vv= arthropod-vectored
434 vertebrate, a= arthropod specific, ba = biting-arthropod (ambiguous state), v = vertebrate
435 (ambiguous state) and ap=plant-sap-feeding-arthropod (ambiguous state). Names in bold and
436 underlined are viruses discovered in this study. The tree is rooted with the Chuvirus clade (root
437 collapsed) as identified as an outgroup in [13] but we note this gives the same result as midpoint
438 and the molecular clock rooting. Nodes labelled with question marks represent nodes with aLRT
439 (approximate likelihood ratio test) statistical support values less than 0.75. Scale bar shows
440 number of amino-acid substitutions per site.
441
442 To test the ability of these reconstructions to correctly predict host associations we
443 randomly selected 10 viruses with known host associations, and assigned them to an
444 ambiguous state. This analysis correctly returned the true host association for all 10 of
445 the randomly chosen viruses with strong posterior support (mean support 0.97, range
446 0.92-1.00).
447
448 *Ancestral host associations and host-switches*
449
450 Viral sequences from arthropods, vertebrates and plants form distinct clusters in the
451 phylogeny (Figure 2). To quantify this genetic structure we calculated a variant of the F_{st}
452 statistic between the sequences of viruses from different groups of hosts. There is
453 strong evidence of genetic differentiation between the sequences from different host
454 groups (P<0.001, Figure S1 http://dx.doi.org/10.6084/m9.figshare.1495351).
455
456 Our Bayesian analysis allowed us to infer the ancestral host association of 178 of 188 of
457 the internal nodes on the phylogenetic tree (support >0.95). A striking pattern that
458 emerged is that switches between major groups of hosts have occurred rarely during
459 the evolution of the rhabdoviruses (Figure 2). There are a few rare transitions on
460 terminal branches (Santa Barbara virus and the virus identified from the plant *Humulus*
461 *lupulus*) but these could represent errors in the host assignment (e.g. cross-species
462 contamination) as well as recent host shifts. Our analysis allows us to estimate the
463 number of times the viruses have switched between major host groups while accounting
464 for uncertainty about ancestral states, and the tree topology and root. There were only
465 two types of host-switch that we found evidence for. Our best estimate of the number of
466 switches from being an arthropod-vectored vertebrate virus to being arthropod specific
467 was three, although one is a jump on a terminal branch may represent contamination
468 (best modal estimate = 3, median = 3.1, CI's= 1.9–6.1). Similarly, we estimate there have
469 been three transitions from being an arthropod-vectored vertebrate virus to a
470 vertebrate-specific virus (best modal estimate = 3, median = 3.1, CI's= 2.9–5.3). We
471 could not determine the direction of the host shifts into the other host groups.
472

473 Vertebrate–specific viruses have arisen once in the lyssaviruses clade [3], as well as at
474 least once in fish dimarhabdoviruses (in one of the fish-infecting clades it is unclear if it
475 is vertebrate-specific or vector-borne from our reconstructions). There has also likely
476 been a single transition to being arthropod-vectored vertebrate viruses in the
477 dimarhabodovirus clade.
478
479 Insect-vectored plant viruses have arisen once in the cyto- and nucleo- rhabdoviruses,
480 although the ancestral state of these viruses is uncertain. A single virus identified from
481 the hop plant *Humulus lupulus* appears to fall within the dimarhabdovirus clade.
482 However, this may be because the plant was contaminated with insect matter, as the
483 same RNA-seq dataset contains *COI* sequences with high similarity to thrips.
484
485 There are two large clades of arthropod-specific viruses. The first is a sister group to the
486 large plant virus clade. This novel group of largely insect-associated viruses are
487 associated with a broad range of insects, including flies, butterflies, moths, ants, thrips,
488 bedbugs, fleas, mosquitoes, water striders and leafhoppers. The mode of transmission
489 and biology of these viruses is yet to be examined. The second clade of insect-associated
490 viruses is the sigma virus clade [11, 12, 19, 45]. These are derived from vector-borne
491 dimarhabdoviruses that have lost their vertebrate host and become vertically
492 transmitted viruses of insects [10]. They are common in Drosophilidae, and our results
493 suggest that they may be widespread throughout the Diptera, with occurrences in the
494 Tephritid fruit fly *Ceratitis capitata*, the stable fly *Muscina stabulans,* several divergent
495 viruses in the housefly *Musca domestica* and louse flies removed from the skin of bats.
496 For the first time we have found sigma-like viruses outside of the Diptera, with two
497 Lepidoptera associated viruses and a virus from an aphid/parasitoid wasp. All of the
498 sigma viruses characterised to date have been vertically transmitted [10], but some of
499 the recently described viruses may be transmitted horizontally – it has been speculated
500 that the viruses from louse flies may infect bats [49] and Shayang Fly Virus 2 has been
501 reported in two fly species [13] (although contamination could also explain this
502 result).Drosophila sigma virus genomes are characterised by an additional X gene
503 between the P and M genes [45]. Interestingly the two louse fly viruses with complete
504 genomes, Wuhan insect virus 7 from an aphid/parasitoid and *Pararge aegeria*
505 rhabdovirus do not have a putative X gene. The first sigma virus was discovered in
506 *Drosophila melanogaster* in 1937 [50], in the last few years related sigma viruses have
507 been found in other *Drosophila* species and a Muscid fly [10-12, 45]. Here we have found
508 sigma-like viruses in a diverse array of Diptera species, as well as other insect orders.
509 Overall, our results suggest sigma-like viruses may be associated with a wide array of
510 insect species.
511
512 Within the arthropod-associated viruses (the most sampled host group) it is common to
513 find closely related viruses in closely related hosts (Figure 1). This is reflected in a
514 positive correlation between the evolutionary distance between the viruses and the
515 evolutionary distance between their arthropod hosts (Pearson's correlation=0.36, 95%
516 CI's=0.34-0.38, P<0.001, Figure 3 and Figure S2
517 http://dx.doi.org/10.6084/m9.figshare.1495351). Since the virus phylogeny is
518 incongruent with that of the respective hosts, this suggests rhabdoviruses preferentially
519 host shift between closely related species [51, 52].

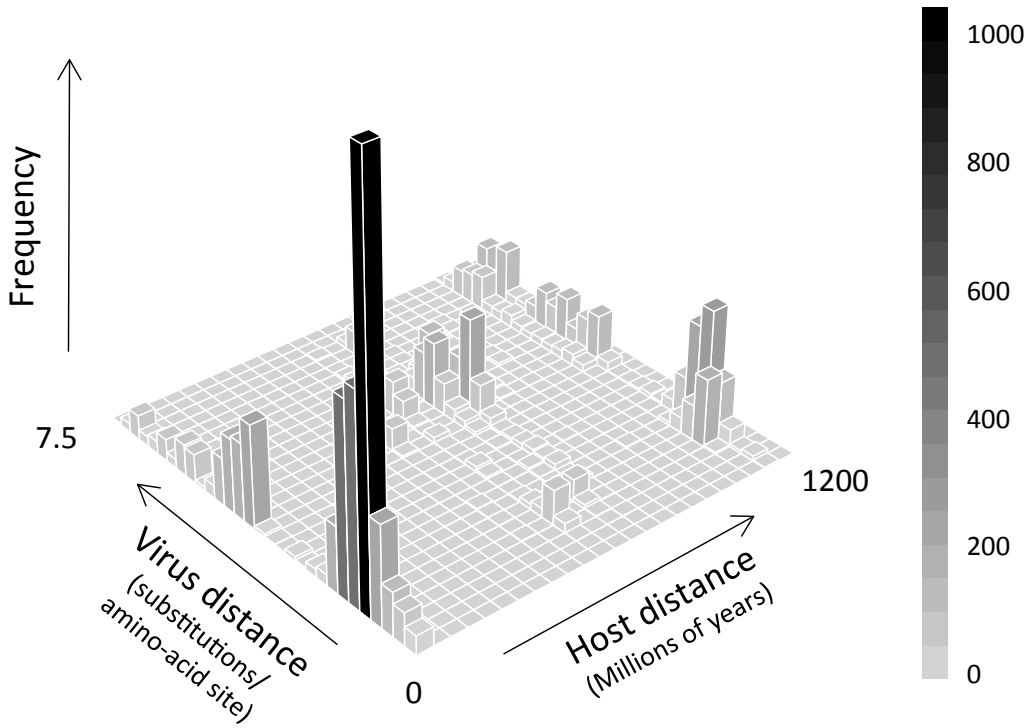


Figure 3. The relationship between the evolutionary distance between viruses and the evolutionary distance between their arthropod hosts (categorised by genus). Closely related viruses tend to be found in closely related hosts. Permutation tests find a significant positive correlation (correlation=0.36, 95% CI's=0.34-0.38, $P<0.001$) between host and virus evolutionary distance (see Figure S2).

We also find viruses clustering on the phylogeny based on the ecosystem of their hosts; there is strong evidence of genetic differentiation between viruses from terrestrial and aquatic hosts ($P=0.007$, Figure S3 http://dx.doi.org/10.6084/m9.figshare.1495351). There has been one shift from terrestrial to aquatic hosts during the evolution of the basal novirhabdoviruses, which have a wide host range in fish. There have been other terrestrial to aquatic shifts in the dimarhabdoviruses: in the clades of fish and cetacean viruses and the clade of viruses from sea-lice.

Discussion

Viruses are ubiquitous in nature and recent developments in high-throughput sequencing technology have led to the discovery and sequencing of a large number of novel viruses in arthropods [13, 14]. Here we have identified 43 novel virus-like sequences, from our own RNA-seq data and public sequence repositories. Of these, 32 were rhabdoviruses, and 26 of these were from arthropods. Using these sequences we have produced the most extensive phylogeny of the *Rhabdoviridae* to date, including a total of 195 virus sequences.

In most cases we know nothing about the biology of the viruses beyond the host they were isolated from, but our analysis provides a powerful way to predict which are vector-borne viruses and which are specific to vertebrates or arthropods. We have

549 identified a large number of new likely vector-borne viruses – of 85 rhabdoviruses
550 identified from vertebrates or biting insects we predict that 76 are arthropod-borne
551 viruses of vertebrates (arboviruses). Re-running the analyses with known host-group
552 excluded for a small number of taxa suggests that this analysis is robust and can
553 correctly predice host-group. Along with the known arboviruses, this suggests the
554 majority of known rhabdoviruses are arboviruses, and all of these fall in a single clade
555 known as the dimarhabdoviruses. In addition to the arboviruses, we also identified two
556 clades of likely insect-specific viruses associated with a wide range of species,
557 suggesting rhabdoviruses may be associated with a wide array of arthropods
558
559 We found that shifts between distantly related hosts are rare in the rhabdoviruses,
560 which is perhaps unsurprising as both rhabdoviruses of vertebrates (rabies virus in
561 bats) and invertebrates (sigma viruses in Drosophildae) show a declining ability to
562 infect hosts more distantly related to their natural host [52-54]. It is thought that sigma
563 viruses may sometimes jump into distantly related but highly susceptible species [51,
564 52, 55], but our results suggest that this rarely happens between major groups such as
565 vertebrates and arthropods. It is nonetheless surprising that arthropod-specific viruses
566 have arisen rarely, as one might naively assume that there would be fewer constraints
567 on vector-borne viruses losing one of their hosts. Within the major clades, closely
568 related viruses often infect closely related hosts (Figure 2). For example, within the
569 dimarhabdoviruses viruses identified from mosquitoes, ticks, *Drosophila*, Muscid flies,
570 Lepidoptera and sea-lice all tend to cluster together (Figure 2B). However, it is also
571 clear that the virus phylogeny does not mirror the host phylogeny, and our data on the
572 clustering of hosts across the virus phylogeny suggest that following major transitions
573 between distantly related host taxa, viruses preferentially shift between more closely
574 related species (Figures 3, S1 and S2) in the same environment (Figure S3).
575
576 There has been a near four-fold increase in the number of recorded rhabdovirus
577 sequences in the last five years. In part this may be due to the falling cost of sequencing
578 transcriptomes [56], and initiatives to sequence large numbers of insect and other
579 arthropods [35]. The use of high-throughput sequencing technologies should reduce the
580 likelihood of sampling biases associated with PCR based discovery where people look
581 for similar viruses in related hosts. This suggests that the pattern of viruses forming
582 clades based on the host taxa they infect is likely to be robust. However, these efforts are
583 disproportionately targeted at arthropods, and it is possible that there may be a great
584 undiscovered diversity of viruses in other organisms [57].
585
586 Rhabdoviruses infect a diverse range of host species, including a large number of
587 arthropods. Our search has unearthed a large number of novel rhabdovirus genomes,
588 suggesting that we are only just beginning to uncover the diversity of these viruses. The
589 host use by these viruses has been highly conserved during their evolution, which
590 provides a powerful tool to identify previously unknown arboviruses. Given the large
591 number of viruses currently being discovered through metagenomic studies [13, 58, 59],
592 in the future we will be faced by an increasingly large number of viral sequences with
593 little knowledge of the biology of the virus. Our phylogenetic approach can be extended
594 to predict key biological traits in other groups of novel pathogens where our knowledge
595 is incomplete. However, the rapid evolution of RNA viruses may mean traits may change

over short timescales, and such an approach should complement, and not replace, examining the basic biology of novel viruses.

Acknowledgments

Many thanks to Mike Ritchie for providing the DMonSV infected fly line; Casper Breuker and Melanie Gibbs for PAegRV samples and Philip Leftwich for CCapSV samples. Thanks to everyone who provided fly collections.

Contributions

BL and FMJ conceived and designed the study. BL and JD carried out molecular work. BL, WJP, DJP and DJO carried out bioinformatic analysis. BL, GGRM and JJW carried out phylogenetic analysis. BL and FMJ wrote the manuscript with comments from all other authors. All authors gave final approval for publication.

Funding

BL and FMJ are supported by a NERC grant (NE/L004232/1), a European Research Council grant (281668, DrosophilaInfection), a Junior Research Fellowship from Christ's College, Cambridge (BL). GGRM is supported by an MRC studentship. The metagenomic sequencing of viruses from *D. immigrans, D. tristis* and *S. deflexa* was supported by a Wellcome Trust fellowship (WT085064) to DJO.

References

1. Lipkin W.I., Anthony S.J. 2015 Virus hunting. *Virology* **479-480C**, 194-199. (doi:10.1016/j.virol.2015.02.006).
2. Liu S., Vijayendran D., Bonning B.C. 2011 Next generation sequencing technologies for insect virus discovery. *Viruses* **3**(10), 1849-1869. (doi:10.3390/v3101849).
3. Dietzgen R.G., Kuzmin I.V. 2012 *Rhabdoviruses: Molecular Taxonomy, Evolution, Genomics, Ecology, Host-Vector Interactions, Cytopathology and Control*. Norfolk, UK, Caister Academic Press.
4. Bourhy H., Cowley J.A., Larrous F., Holmes E.C., Walker P.J. 2005 Phylogenetic relationships among rhabdoviruses inferred using the L polymerase gene. *Journal of General Virology* **86**, 2849-2858.
5. Hampson K., Coudeville L., Lembo T., Sambo M., Kieffer A., Attlan M., Barrat J., Blanton J.D., Briggs D.J., Cleaveland S., et al. 2015 Estimating the global burden of endemic canine rabies. *PLoS Negl Trop Dis* **9**(4), e0003709. (doi:10.1371/journal.pntd.0003709).
6. Walker P.J., Blasdell K.R., Joubert D.A. 2012 Ephemeroviruses: Athropod-borne Rhabdoviruses of Ruminants, with Large, Complex Genomes. In *Rhabdoviruses: Molecular Taxonomy, Evolution, Genomics, Ecology, Host-Vector Interactions, Cytopathology and Control* (eds. Dietzgen R.G., Kuzmin I.V.), pp. 59-88. Norfolk, UK, Caister Academic Press.
7. Hogenhout S.A., Redinbaugh M.G., Ammar E.D. 2003 Plant and animal rhabdovirus host range: a bug's view. *Trends in Microbiology* **11**(6), 264-271. (doi:Doi 10.1016/S0966-842x(03)00120-3).

646 8. Ahne W., Bjorklund H.V., Essbauer S., Fijan N., Kurath G., Winton J.R. 2002 Spring
647 viremia of carp (SVC). *Diseases of Aquatic Organisms* **52**, 261-272.
648 9. Pfeilputzien C. 1978 EXPERIMENTAL TRANSMISSION OF SPRING VIREMIA OF
649 CARP THROUGH CARP LICE (ARGULUS-FOLIACEUS). *Zentralblatt Fur Veterinarmedizin*
650 *Reihe B-Journal of Veterinary Medicine Series B-Infectious Diseases Immunology Food*
651 *Hygiene Veterinary Public Health* **25**(4), 319-323.
652 10. Longdon B., Jiggins F.M. 2012 Vertically transmitted viral endosymbionts of
653 insects: do sigma viruses walk alone? *Proc Biol Sci* **279**(1744), 3889-3898.
654 (doi:10.1098/rspb.2012.1208).
655 11. Longdon B., Wilfert L., Obbard D.J., Jiggins F.M. 2011 Rhabdoviruses in two
656 species of Drosophila: vertical transmission and a recent sweep. *Genetics* **188**(1), 141-
657 150. (doi:10.1534/genetics.111.127696).
658 12. Longdon B., Wilfert L., Osei-Poku J., Cagney H., Obbard D.J., Jiggins F.M. 2011
659 Host switching by a vertically-transmitted rhabdovirus in Drosophila. *Biology Letters*
660 **7**(5), 747-750. (doi:10.1098/rsbl.2011.0160).
661 13. Li C.X., Shi M., Tian J.H., Lin X.D., Kang Y.J., Chen L.J., Qin X.C., Xu J., Holmes E.C.,
662 Zhang Y.Z. 2015 Unprecedented genomic diversity of RNA viruses in arthropods reveals
663 the ancestry of negative-sense RNA viruses. *eLife* **4**. (doi:10.7554/eLife.05378).
664 14. Walker P.J., Firth C., Widen S.G., Blasdell K.R., Guzman H., Wood T.G., Paradkar
665 P.N., Holmes E.C., Tesh R.B., Vasilakis N. 2015 Evolution of genome size and complexity
666 in the rhabdoviridae. *PLoS Pathog* **11**(2), e1004664.
667 (doi:10.1371/journal.ppat.1004664).
668 15. Ballinger M.J., Bruenn J.A., Taylor D.J. 2012 Phylogeny, integration and
669 expression of sigma virus-like genes in Drosophila. *Mol Phylogenet Evol* **65**(1), 251-258.
670 (doi:10.1016/j.ympev.2012.06.008).
671 16. Fort P., Albertini A., Van-Hua A., Berthomieu A., Roche S., Delsuc F., Pasteur N.,
672 Capy P., Gaudin Y., Weill M. 2011 Fossil Rhabdoviral Sequences Integrated into
673 Arthropod Genomes: Ontogeny, Evolution, and Potential Functionality. *Mol Biol Evol*.
674 (doi:10.1093/molbev/msr226).
675 17. Katzourakis A., Gifford R.J. 2010 Endogenous Viral Elements in Animal Genomes.
676 *Plos Genet* **6**(11), e1001191. (doi:10.1371/journal.pgen.1001191).
677 18. Aiewsakun P., Katzourakis A. 2015 Endogenous viruses: Connecting recent and
678 ancient viral evolution. *Virology* **479-480C**, 26-37. (doi:10.1016/j.virol.2015.02.011).
679 19. Longdon B., Wilfert L., Jiggins F.M. 2012 *The Sigma Viruses of Drosophila*, Caister
680 Academic Press.
681 20. Rosen L. 1980 Carbon-dioxide sensitivity in mosquitos infected with sigma,
682 vesicular stomatitis, and other rhabdoviruses. *Science* **207**(4434), 989-991.
683 21. Shroyer D.A., Rosen L. 1983 Extrachromosomal-inheritance of carbon-dioxide
684 sensitivity in the mosquito culex-quinquefasciatus. *Genetics* **104**(4), 649-659.
685 22. Parker D.J., Vesala L., Ritchie M.G., Laiho A., Hoikkala A., Kankare M. 2015 How
686 consistent are the transcriptome changes associated with cold acclimation in two
687 species of the Drosophila virilis group? *Heredity (Edinb)*. (doi:10.1038/hdy.2015.6).
688 23. van Mierlo J.T., Overheul G.J., Obadia B., van Cleef K.W., Webster C.L., Saleh M.C.,
689 Obbard D.J., van Rij R.P. 2014 Novel Drosophila viruses encode host-specific suppressors
690 of RNAi. *PLoS Pathog* **10**(7), e1004256. (doi:10.1371/journal.ppat.1004256).
691 24. Katoh K., Standley D.M. 2013 MAFFT multiple sequence alignment software
692 version 7: improvements in performance and usability. *Mol Biol Evol* **30**(4), 772-780.
693 (doi:10.1093/molbev/mst010).
694 25. Capella-Gutierrez S., Silla-Martinez J.M., Gabaldon T. 2009 trimAl: a tool for
695 automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*
696 **25**(15), 1972-1973. (doi:10.1093/bioinformatics/btp348).
697 26. Talavera G., Castresana J. 2007 Improvement of phylogenies after removing
698 divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol*
699 **56**(4), 564-577. (doi:Doi 10.1080/10635150701472164).

700 27. Guindon S., Dufayard J.F., Lefort V., Anisimova M., Hordijk W., Gascuel O. 2010
701 New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing
702 the Performance of PhyML 3.0. *Syst Biol* **59**(3), 307-321. (doi:Doi
703 10.1093/Sysbio/Syq010).
704 28. Le S.Q., Gascuel O. 2008 An improved general amino acid replacement matrix.
705 *Molecular Biology and Evolution* **25**(7), 1307-1320. (doi:Doi 10.1093/Molbev/Msn067).
706 29. Anisimova M., Gascuel O. 2006 Approximate likelihood-ratio test for branches: A
707 fast, accurate, and powerful alternative. *Syst Biol* **55**(4), 539-552.
708 (doi:10.1080/10635150600755453).
709 30. Rambaut A. 2011 FigTree. (v1.3 ed.
710 31. Hudson R.R., Slatkin M., Maddison W.P. 1992 Estimation of levels of gene flow
711 from DNA sequence data. *Genetics* **132**(2), 583-589.
712 32. Bhatia G., Patterson N., Sankararaman S., Price A.L. 2013 Estimating and
713 interpreting FST: the impact of rare variants. *Genome Res* **23**(9), 1514-1521.
714 (doi:10.1101/gr.154831.113).
715 33. Hommola K., Smith J.E., Qiu Y., Gilks W.R. 2009 A permutation test of host-
716 parasite cospeciation. *Mol Biol Evol* **26**(7), 1457-1468. (doi:10.1093/molbev/msp062).
717 34. Jeyaprakash A., Hoy M.A. 2009 First divergence time estimate of spiders,
718 scorpions, mites and ticks (subphylum: Chelicerata) inferred from mitochondrial
719 phylogeny. *Experimental & applied acarology* **47**(1), 1-18. (doi:10.1007/s10493-008-
720 9203-5).
721 35. Misof B., Liu S., Meusemann K., Peters R.S., Donath A., Mayer C., Frandsen P.B.,
722 Ware J., Flouri T., Beutel R.G., et al. 2014 Phylogenomics resolves the timing and pattern
723 of insect evolution. *Science* **346**(6210), 763-767. (doi:10.1126/science.1257570).
724 36. Bootsma R., Dekinkelin P., Leberre M. 1975 Transmission Experiments with Pike
725 Fry (Esox-Lucius L) Rhabdovirus. *J Fish Biol* **7**(2), 269-276. (doi:Doi 10.1111/J.1095-
726 8649.1975.Tb04599.X).
727 37. Dorson M., Dekinkelin P., Torchy C., Monge D. 1987 SUSCEPTIBILITY OF PIKE
728 (ESOX-LUCIUS) TO DIFFERENT SALMONID VIRUSES (IPN, VHS, IHN) AND TO THE
729 PERCH RHABDOVIRUS. *BULLETIN FRANCAIS DE LA PECHE ET DE LA PISCICULTURE*
730 (307), 91-101.
731 38. Haenen O., Davidse A. 1993 Comparative pathogenicity of two strains of pike fry
732 rhabdovirus and spring viremia of carp virus for young roach, common carp, grass carp
733 and rainbow trout. *Diseases of Aquatic Organisms* **15**(2), 87-92.
734 39. Okland A.L., Nylund A., Overgard A.C., Blindheim S., Watanabe K., Grotmol S.,
735 Arnesen C.E., Plarre H. 2014 Genomic characterization and phylogenetic position of two
736 new species in Rhabdoviridae infecting the parasitic copepod, salmon louse
737 (Lepeophtheirus salmonis). *Plos One* **9**(11), e112517.
738 (doi:10.1371/journal.pone.0112517).
739 40. Drummond A.J., Suchard M.A., Xie D., Rambaut A. 2012 Bayesian phylogenetics
740 with BEAUti and the BEAST 1.7. *Mol Biol Evol* **29**(8), 1969-1973.
741 (doi:10.1093/molbev/mss075).
742 41. Weinert L.A., Welch J.J., Suchard M.A., Lemey P., Rambaut A., Fitzgerald J.R. 2012
743 Molecular dating of human-to-bovid host jumps by Staphylococcus aureus reveals an
744 association with the spread of domestication. *Biol Lett* **8**(5), 829-832.
745 (doi:10.1098/rsbl.2012.0290).
746 42. Drummond A.J., Ho S.Y., Phillips M.J., Rambaut A. 2006 Relaxed phylogenetics
747 and dating with confidence. *PLoS Biol* **4**(5), e88. (doi:10.1371/journal.pbio.0040088).
748 43. Minin V.N., Suchard M.A. 2008 Counting labeled transitions in continuous-time
749 Markov models of evolution. *Journal of mathematical biology* **56**(3), 391-412.
750 (doi:10.1007/s00285-007-0120-8).
751 44. Rambaut A., Drummond A.J. 2007. *Tracer v16, Available from*
752 http://beast.bio.ed.ac.uk/Tracer

753 45. Longdon B., Obbard D.J., Jiggins F.M. 2010 Sigma viruses from three species of
754 Drosophila form a major new clade in the rhabdovirus phylogeny. *Proceedings of the*
755 *Royal Society B* **277**, 35-44. (doi:10.1098/rspb.2009.1472).
756 46. Longdon B., Walker P.J. 2011 ICTV sigmavirus species and genus proposal. (
757 47. Walker P.J., Dietzgen R.G., Joubert D.A., Blasdell K.R. 2011 Rhabdovirus accessory
758 genes. *Virus Res* **162**(1-2), 110-125. (doi:10.1016/j.virusres.2011.09.004).
759 48. Chiba S., Kondo H., Tani A., Saisho D., Sakamoto W., Kanematsu S., Suzuki N. 2011
760 Widespread Endogenization of Genome Sequences of Non-Retroviral RNA Viruses into
761 Plant Genomes. *Plos Pathogens* **7**(7). (doi:Artn E1002146
762 Doi 10.1371/Journal.Ppat.1002146).
763 49. Aznar-Lopez C., Vazquez-Moron S., Marston D.A., Juste J., Ibanez C., Berciano J.M.,
764 Salsamendi E., Aihartza J., Banyard A.C., McElhinney L., et al. 2013 Detection of
765 rhabdovirus viral RNA in oropharyngeal swabs and ectoparasites of Spanish bats. *J Gen*
766 *Virol* **94**(Pt 1), 69-75. (doi:10.1099/vir.0.046490-0).
767 50. L'Heritier P.H., Teissier G. 1937 Une anomalie physiologique héréditaire chez la
768 Drosophile. *CR Acad Sci Paris* **231**, 192-194.
769 51. Longdon B., Brockhurst M.A., Russell C.A., Welch J.J., Jiggins F.M. 2014 The
770 Evolution and Genetics of Virus Host Shifts. *PLoS Pathog* **10**(11), e1004395.
771 (doi:10.1371/journal.ppat.1004395).
772 52. Longdon B., Hadfield J.D., Webster C.L., Obbard D.J., Jiggins F.M. 2011 Host
773 phylogeny determines viral persistence and replication in novel hosts. *PLoS Pathogens*
774 **7**((9)), e1002260. (doi:10.1371/journal.ppat.1002260).
775 53. Faria N.R., Suchard M.A., Rambaut A., Streicker D.G., Lemey P. 2013
776 Simultaneously reconstructing viral cross-species transmission history and identifying
777 the underlying constraints. *Philos Trans R Soc Lond B Biol Sci* **368**(1614), 20120196.
778 (doi:10.1098/rstb.2012.0196).
779 54. Streicker D.G., Turmelle A.S., Vonhof M.J., Kuzmin I.V., McCracken G.F., Rupprecht
780 C.E. 2010 Host Phylogeny Constrains Cross-Species Emergence and Establishment of
781 Rabies Virus in Bats. *Science* **329**(5992), 676-679. (doi:10.1126/science.1188836).
782 55. Longdon B., Hadfield J.D., Day J.P., Smith S.C., McGonigle J.E., Cogni R., Cao C.,
783 Jiggins F.M. 2015 The Causes and Consequences of Changes in Virulence following
784 Pathogen Host Shifts. *PLoS Pathog* **11**(3), e1004728.
785 (doi:10.1371/journal.ppat.1004728).
786 56. Wang Z., Gerstein M., Snyder M. 2009 RNA-Seq: a revolutionary tool for
787 transcriptomics. *Nat Rev Genet* **10**(1), 57-63. (doi:10.1038/nrg2484).
788 57. Dudas G., Obbard D.J. 2015 Are arthropods at the heart of virus evolution? *eLife*
789 **4**.
790 58. Aguiar E.R., Olmo R.P., Paro S., Ferreira F.V., de Faria I.J., Todjro Y.M., Lobo F.P.,
791 Kroon E.G., Meignin C., Gatherer D., et al. 2015 Sequence-independent characterization
792 of viruses based on the pattern of viral small RNAs produced by the host. *Nucleic Acids*
793 *Res*. (doi:10.1093/nar/gkv587).
794 59. Webster C.L., Waldron F.M., Robertson S., Crowson D., Ferrai G., Quintana J.F.,
795 Brouqui J.M., Bayne E.H., Longdon B., Buck A.H., et al. 2015 The discovery, distribution
796 and evolution of viruses associated with Drosophila melanogaster. *PLOS Biology* **13(7):**
797 **e1002210.**
798
799
800 **All data has been made available in public repositories:**
801
802 NCBI Sequence Read Archive Data: SRP057824
803 Data S1, sample information: http://dx.doi.org/10.6084/m9.figshare.1425432
804 Data S2, virus ID, Genbank accessions and host information:
805 http://dx.doi.org/10.6084/m9.figshare.1425419

806 Figures S1-S3: http://dx.doi.org/10.6084/m9.figshare.1495351
807 L gene sequences fasta: http://dx.doi.org/10.6084/m9.figshare.1425067
808 TrimAl alignment fasta: http://dx.doi.org/10.6084/m9.figshare.1425069
809 Gblocks alignment fasta: http://dx.doi.org/10.6084/m9.figshare.1425068
810 Phylogenetic tree Gblocks alignment: http://dx.doi.org/10.6084/m9.figshare.1425083
811 Phylogenetic tree TrimAl alignment: http://dx.doi.org/10.6084/m9.figshare.1425082
812 BEAST alignment fasta: http://dx.doi.org/10.6084/m9.figshare.1425431
813 BEAUti xml file: http://dx.doi.org/10.6084/m9.figshare.1431922
814 Bayesian analysis tree: http://dx.doi.org/10.6084/m9.figshare.1425436
815 Tables S1-5. List of newly discovered viruses:
816 http://dx.doi.org/10.6084/m9.figshare.1502665
817
818