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 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]. Insects are not just mechanical vectors as rhabdoviruses replicate upon infection of insect 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

new likely vector-borne viruses and to reconstruct the evolutionary history of this diverse group of viruses.

Methods

Discovery of new rhabdoviruses by RNA sequencing

Diptera (flies, mostly Drosophilidae) were collected in the field from Spain, USA, Kenya, France, Ghana and the UK (Data S1: http://dx.doi.org/10.6084/m9.figshare.1425432). Infection with rhabdoviruses can cause *Drosophila* and other insects to become paralysed after exposure to CO_2 [19-21], so we enriched our sample for infected individuals by exposing them to CO_2 at 12°C for 15 mins, only retaining individuals that showed symptoms of paralysis 30mins later. We extracted RNA from 79 individual insects (details in Data S1 http://dx.doi.org/10.6084/m9.figshare.1425432) using Trizol reagent (Invitrogen) and combined the extracts into two pools (retaining non-pooled individual RNA samples). RNA was then rRNA depleted with the Ribo-Zero Gold kit (epicenter, USA) and used to construct Truseq total RNA libraries (Illumina). Libraries were constructed and sequenced by BGI (Hong Kong) on an Illumina Hi-Seq 2500 (one lane, 100bp paired end reads, generating ~175 million reads). Sequences were quality trimmed with Trimmomatic (v3); Illumina adapters were clipped, bases were removed from the beginning and end of reads if quality dropped below a threshold, sequences were trimmed if the average quality within a window fell below a threshold and reads less than 20 base pairs in length were removed. We *de novo* assembled the RNA-seq reads with Trinity (release 2013-02-25) using default settings and jaccard clip option for high gene density. The assembly was then searched using tblastn to identify rhabdovirus-like sequences, with known rhabdovirus coding sequences as the query. Contigs with hits were then reciprocally compared to Genbank cDNA and RefSeq nucleotide databases using tblastn and only retained if they hit a virus-like sequence. Raw read data were deposited in the NCBI Sequence Read Archive (SRP057824). Putative viral sequences have been submitted to Genbank (accessions in Tables S1 and S2 http://dx.doi.org/10.6084/m9.figshare.1502665).

As the RNA-seq was performed on pooled samples, we assigned rhabdovirus sequences to individual insects by PCR on individual samples RNA. cDNA was produced using Promega GoScript Reverse Transcriptase and random-hexamer primers, and PCR performed using primers designed using the rhabdovirus sequences. Infected host species were identified by sequencing the mitochondrial gene *COI*. We were unable to identify the host species of the virus from a *Drosophila affinis* sub-group species (sequences appear similar to both *Drosophila affinis* and the closely related Drosophila *athabasca*), despite the addition of further mitochondrial and nuclear sequences to increase confidence. In all cases we confirmed that viruses were only present in cDNA and not in non reverse-transcription (RT) controls (i.e. DNA) by PCR, and so they cannot be integrated into the insect genome (i.e. endogenous virus elements or EVEs [17]). *COI* primers were used as a positive control for the presence of DNA in the non RT template.

We identified sigma virus sequences in RNA-seq data from *Drosophila montana* [22]. We used RT-PCR on an infected fly line to amplify the virus sequence, and carried out additional Sanger sequencing with primers designed using the RNA-seq assembly. Additional virus sequences were identified from an RNA-seq analysis of pools of wild caught *Drosophila*: DImmSV from *Drosophila immigrans* (collection and sequencing described [23]), DTriSV from a pool of *Drosophila tristis* and SDefSV from *Scaptodrosophila deflexa* (both Darren Obbard, unpublished data), accessions in tables S1 and S2 (http://dx.doi.org/10.6084/m9.figshare.1502665).

Discovery of rhabdoviruses in public sequence databases

Rhabdovirus L gene sequences were used to search (tblastn) against expressed sequence tag (EST) and transcriptome shotgun assembly (TSA) databases (NCBI). All hits were reciprocally BLAST searched against Genbank cDNA and RefSeq databases and only retained if they hit a virus-like sequence. We used two approaches to examine whether sequences were present as RNA but not DNA. First, where assemblies of whole-genome shotgun sequences were available, we used BLAST to test whether sequences were integrated into the host genome. Second, for the virus sequences in the butterfly *Pararge aegeria* and the medfly *Ceratitis capitata* we were able to obtain infected samples to confirm the sequences are only present in RNA by performing PCR on both genomic DNA and cDNA as described above (samples kindly provided by Casper Breuker/Melanie Gibbs, and Philip Leftwich respectively)

Phylogenetic analysis

All available rhabdovirus-like sequences were downloaded from Genbank (accessions in Data S2: http://dx.doi.org/10.6084/m9.figshare.1425419). Amino acid sequences for the L gene (encoding the RNA Dependent RNA Polymerase or RDRP) were used to infer the phylogeny (L gene sequences: http://dx.doi.org/10.6084/m9.figshare.1425067), as they contain conserved domains that can be aligned across this diverse group of viruses. Sequences were aligned with MAFFT [24] under default settings and then poorly aligned and divergent sites were removed with either TrimAl (v1.3 strict settings, implemented on Phylemon v2.0 server, alignment: http://dx.doi.org/10.6084/m9.figshare.1425069) [25] or Gblocks (v0.91b selecting smaller final blocks, allowing gap positions and less strict flanking positions to produce a less stringent selection, alignment: http://dx.doi.org/10.6084/m9.figshare.1425068) [26]. These resulted in alignments of 1492 and 829 amino acids respectively.

Phylogenetic trees were inferred using Maximum Likelihood in PhyML (v3.0) [27] using the LG substitution model [28] (preliminary analysis confirmed the results were robust to the amino acid substitution model selected), with a gamma distribution of rate variation with four categories and using a sub-tree pruning and regrafting topology searching algorithm. Branch support was estimated using Approximate Likelihood-Ratio Tests (aLRT) that is reported to outperform bootstrap methods [29]. Figures were created using FIGTREE (v. 1.4) [30].

Analysis of genetic structure between viruses taken from different hosts and ecologies

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

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

where a population is a host category or ecosystem. The significance of this value was tested by permuting the host/ecosystem association of viruses. Since arthropods are the most sampled host, we tested for evidence of genetic structure within the arthropod-associated viruses that would suggest co-divergence with their hosts or preferential host-switching between closely related hosts. We calculated the Pearson correlation coefficient of the evolutionary distances between viruses and the evolutionary distances between their hosts and tested for significance by permutation (as in [33]). We used the patristic distances of our ML tree for the virus data and a time-tree of arthropod genera, using published estimates of divergence dates [34, 35].

Reconstruction of host associations

Viruses were categorised as having one of four types of host association: arthropod-specific, vertebrate-specific, arthropod-vectored plant, or arthropod-vectored vertebrate. However, the host association of some viruses are uncertain when they have been isolated from vertebrates, biting-arthropods or plant-sap-feeding arthropods. Due to limited sampling it was not clear whether viruses isolated from vertebrates were vertebrate specific or arthropod-vectored vertebrate viruses; or whether viruses isolated from biting-arthropods were arthropod specific viruses or arthropod-vectored vertebrate viruses; or if viruses isolated from plant-sap-feeding arthropods were arthropod-specific or arthropod-vectored plant viruses.

We omitted three viruses that were isolated from hosts outside of these four categories (viruses from a nematode, fungus and cnidarian) from our analyses. We classified three of the fish infecting dimarhabdoviruses as vertebrate specific based on the fact they can be transmitted via immersion in water containing virus during experimental conditions [36-38], and the widely held belief amongst the fisheries community that these viruses are not typically vectored [8]. However, there is some evidence these viruses can be transmitted by arthropods (sea lice) in experiments [8, 9] and so we would recommend this be interpreted with some caution. Additionally, although we classified the viruses identified in sea-lice as having biting arthropod hosts, they may be crustacean-specific. The two viruses from *Lepeophtheirus salmonis* do not seem to infect the fish they parasitise and are present in all developmental stages of the lice, suggesting they may be transmitted vertically [39].

We simultaneously estimated both the current and ancestral host associations, and the phylogeny of the viruses, using a Bayesian analysis, implemented in BEAST v1.8 [40, 41]. Since accurate branch lengths are essential for this analysis, we used a subset of the

sites and strains used in the Maximum Likelihood (ML) analysis. We retained 189 taxa; all rhabdoviruses excluding the divergent fish-infecting novirhabdovirus clade and the virus from *Hydra*, as well as the viruses from *Lolium perenne and Conwentzia psociformis*, which had a large number of missing sites. Sequences were trimmed to a conserved region of 414 amino acids where data was recorded for most of these viruses (the Gblocks alignment trimmed further by eye: http://dx.doi.org/10.6084/m9.figshare.1425431). We used the host-association categories described above, which included ambiguous states. To model amino acid evolution we used an LG substitution model with gamma distributed rate variation across sites [28] and an uncorrelated lognormal relaxed clock model of rate variation among lineages [42]. To model the evolution of the host associations we used an asymmetric transition rate matrix (allowing transitions to and from a host association to take place at different rates) and a strict clock model. We also examined how often these viruses jumped between different classes of hosts using reconstructed counts of biologically feasible changes of host association and their HPD confidence intervals (CI) using Markov Jumps [43]. These included switches between arthropod-specific and both arthropod-vectored vertebrate and arthropod-vectored plant states, and between vertebrate specific and arthropod-vectored vertebrate states. We used a constant population size coalescent prior for the relative node ages (using a birth-death prior gave equivalent results) and the BEAUti v1.8 default priors for all other parameters [40] (BEAUti xml http://dx.doi.org/10.6084/m9.figshare.1431922). Convergence was assessed using Tracer v1.6 [44], and a burn-in of 30% was removed prior to the construction of a consensus tree, which included a description of ancestral host associations. High effective sample sizes were achieved for all parameters (>200).

The maximum clade credibility tree estimated (http://dx.doi.org/10.6084/m9.figshare.1425436) for the host association reconstruction was very similar to the independently estimated maximum likelihood phylogeny (http://dx.doi.org/10.6084/m9.figshare.1425083), which made no assumptions about the appropriateness or otherwise of applying a clock model. The minor topological differences may be expected in reconstructions that differ in their assumptions about evolutionary rates. In Figure 2 we have transferred the ancestral state reconstruction from the BEAST tree to the maximum likelihood tree.

Results

Novel rhabdoviruses from RNA-seq

To search for new rhabdoviruses we collected a variety of different species of flies, screened them for CO_2 sensitivity, which is a common symptom of infection, and sequenced total RNA of these flies by RNA-seq. We identified rhabdovirus-like sequences from a *de-novo* assembly by BLAST, and used PCR to identify which samples these sequences came from.

This approach resulted in eleven rhabdovirus-like sequences from nine (possibly ten) species of fly. Seven of these viruses were previously unknown and four had been

reported previously from shorter sequences (Tables S1 and S2 http://dx.doi.org/10.6084/m9.figshare.1502665). The novel viruses were highly divergent from known viruses. Sigma viruses known from other species of *Drosophila* typically have genomes of ~12.5Kb [12, 45], and six of our sequences were approximately this size, suggesting they are near-complete genomes. None of the viruses discovered in our RNA-seq data were integrated into the host genome (see Methods for details).

To investigate the putative gene content of the viruses, we predicted putative genes based on open reading frames (ORFs). For the viruses with apparently complete genomes (Figure 1), we found that those from *Drosophila ananassae, Drosophila affinis, Drosophila immigrans* and *Drosophila sturtvanti* contained ORFs corresponding to the five core genes found across all rhabdoviruses, with an additional ORF between the P and M genes. This is the location of the X gene found in sigma viruses, and in three of the four novel viruses it showed BLAST sequence similarity to the X gene of sigma viruses. The virus from *Drosophila busckii* did not contain an additional ORF between the P and M genes, but instead contained an ORF between the G and L gene. The high mutation rate of RNA viruses mean that such ORFs are highly likely to be functional and retained by selection, as non-function ORFs would quickly be degraded.

Using the phylogeny described below, we have classified our newly discovered viruses as either sigma viruses, rhabdoviruses or other viruses, and named them after the host species they were identified from (Figure 1) [46]. We also found one other novel mononegavirales-like sequence from *Drosophila unispina* that groups with a recently discovered clade of arthropod associated viruses (Nyamivirus clade [13], see Table S5 and the full phylogeny: http://dx.doi.org/10.6084/m9.figshare.1425083), as well as five other RNA viruses from various families (data not shown), confirming our approach can detect a wide range of divergent viruses.

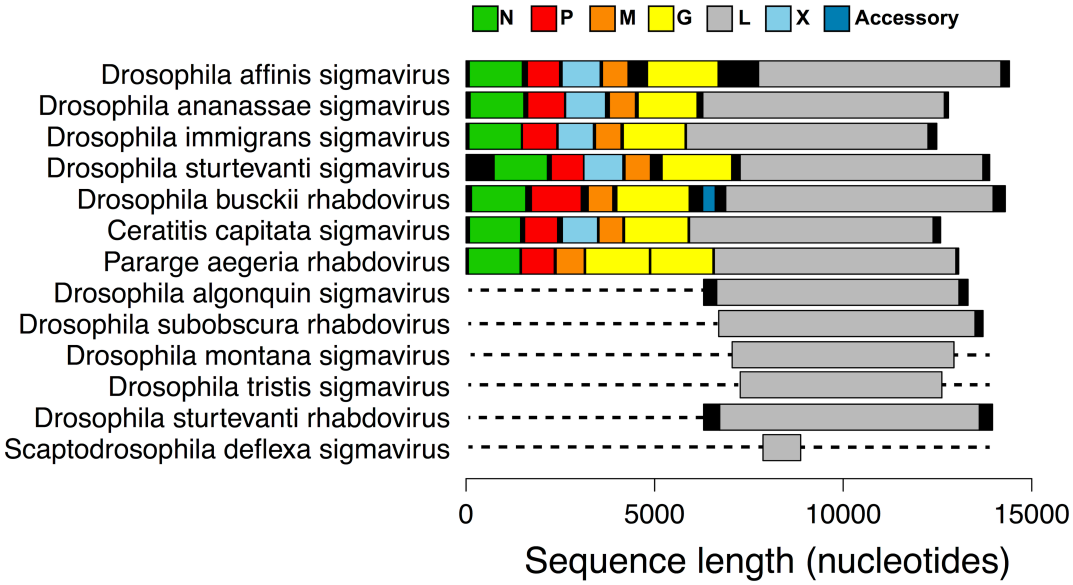


Figure 1. Genome organization of newly discovered viruses.
Putative genes are shown in colour, non-coding regions are shown in black. ORFs were designated as the first start codon following the transcription termination sequence (7 U's) of the

previous ORF to the first stop codon. Dotted lines represent parts of the genome not sequenced. These viruses were either from our own RNA-seq data, or were first found in in public databases and key features verified by PCR and Sanger sequencing. Rhabdovirus genomes are typically ~11-13kb long and contain five core genes 3'-N-P-M-G-L-5' [3]. However, a number of groups of rhabdoviruses contain additional accessory genes and can be up to ~16kb long [14, 47].

New rhabdoviruses from public databases

We identified a further 26 novel rhabdovirus-like sequences by searching public databases of assembled RNA-seq data with BLAST. These included 19 viruses from arthropods (Fleas, Crustacea, Lepidoptera, Diptera), one from a Cnidarian (*Hydra*) and 5 from plants (Table S3). Of these viruses, 19 had sufficient amounts of coding sequence (>1000bp) to include in the phylogenetic analysis (Table S3), whilst the remainder were too short (Table S4).

Four viruses from databases had near-complete genomes based on their size. These were from the moth *Triodia sylvina*, the house fly *Musca domestica* (99% nucleotide identity to Wuhan house fly virus 2 [13]), the butterfly *Pararge aegeria* and the medfly *Ceratitis capitata*, all of which contain ORFs corresponding to the five core rhabdovirus genes. The sequence from *C. capitata* had an additional ORF between the P and M genes with BLAST sequence similarity to the X gene in sigma viruses. There were several unusual sequences. Firstly, in the virus from *P. aegeria* there appear to be two full-length glycoprotein ORFs between the M and L genes (we confirmed by Sanger sequencing that both exist and the stop codon between the two genes was not an error). Secondly, the *Agave tequilana* transcriptome contained a L gene ORF on a contig that was the length of a typical rhabdovirus genome but did not appear to contain typical gene content, suggesting it has very atypical genome organization, or has been misassembled, or is integrated into its host plant genome [48]. Finally, the virus from *Hydra magnipapillata* contained six predicted genes, but the L gene (RDRP) ORF was unusually long. Some of the viruses we detected may be EVEs inserted into the host genome and subsequently expressed [18]. For example, this is likely the case for the sequence from the silkworm *Bombyx mori* that we also found in the silkworm genome, and the L gene sequence from *Spodoptera exigua* that contains stop codons. Under the assumption that viruses integrated into host genomes once infected those hosts, this does not affect our conclusions below about the host range of these viruses [15-17]. We also found nine other novel mononegavirale-like sequences that group with recently discovered clades of insect viruses [13] (see Table S5 and http://dx.doi.org/10.6084/m9.figshare.1425083).

Rhabdovirus Phylogeny

To reconstruct the evolution of the *Rhabdoviridae* we have produced the most complete phylogeny of the group to date (Figure 2). We aligned the relatively conserved L gene (RNA Dependant RNA Polymerase) from our newly discovered viruses with sequences of known rhabdoviruses to give an alignment of 195 rhabdoviruses (and 26 other mononegavirales as an outgroup). We reconstructed the phylogeny using different sequence alignments and methodologies, and these all gave qualitatively similar results

with the same major clades being reconstructed (Gblocks: http://dx.doi.org/10.6084/m9.figshare.1425083, TrimAl: http://dx.doi.org/10.6084/m9.figshare.1425082 and BEAST: http://dx.doi.org/10.6084/m9.figshare.1425436). The ML and Bayesian relaxed clock phylogenies were very similar, suggesting our analysis is robust to the assumptions of a relaxed molecular clock. The branching order between the clades in the dimarhabdovirus supergroup was generally poorly supported and differed between the methods and alignments. Eight sequences that we discovered were not included in this analysis as they were considered too short, but their closest BLAST hits are listed in Table S4 (http://dx.doi.org/10.6084/m9.figshare.1502665).

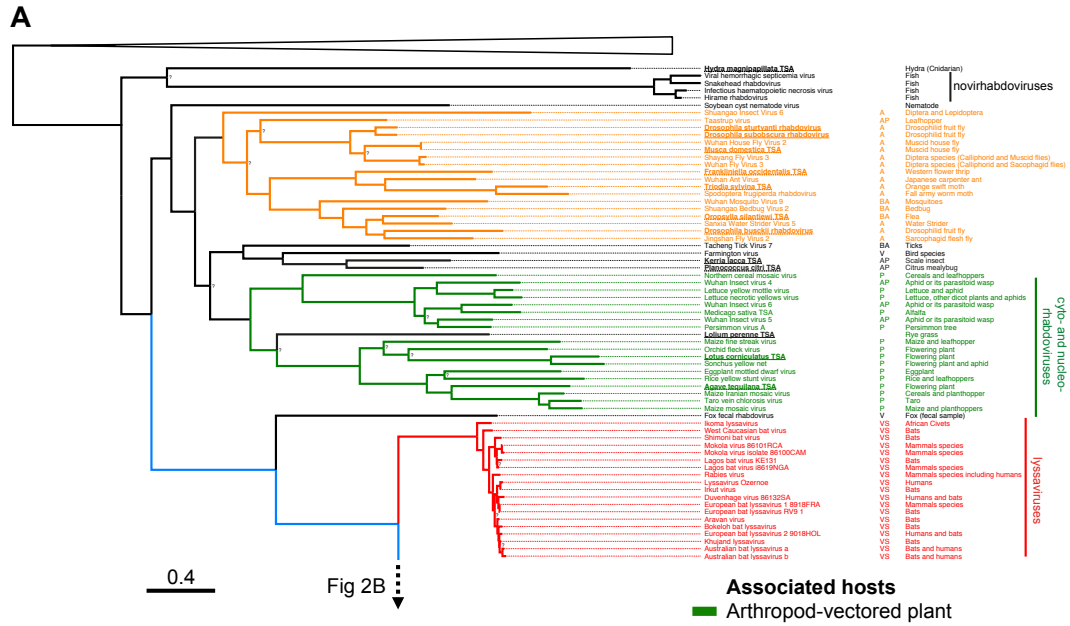
We recovered all of the major clades described previously (Figure 2), and found that the majority of known rhabdoviruses belong to the dimarhabdovirus clade (Figure 2b). The RNA-seq viruses from *Drosophila* fall into either the sigma virus clade (Figure 2b) or the arthropod clade sister to the cyto- and nucleo- rhabdoviruses (Figure 2a). The viruses from sequence databases are diverse, coming from almost all of the major clades with the exception of the lyssaviruses.

Predicted host associations of viruses

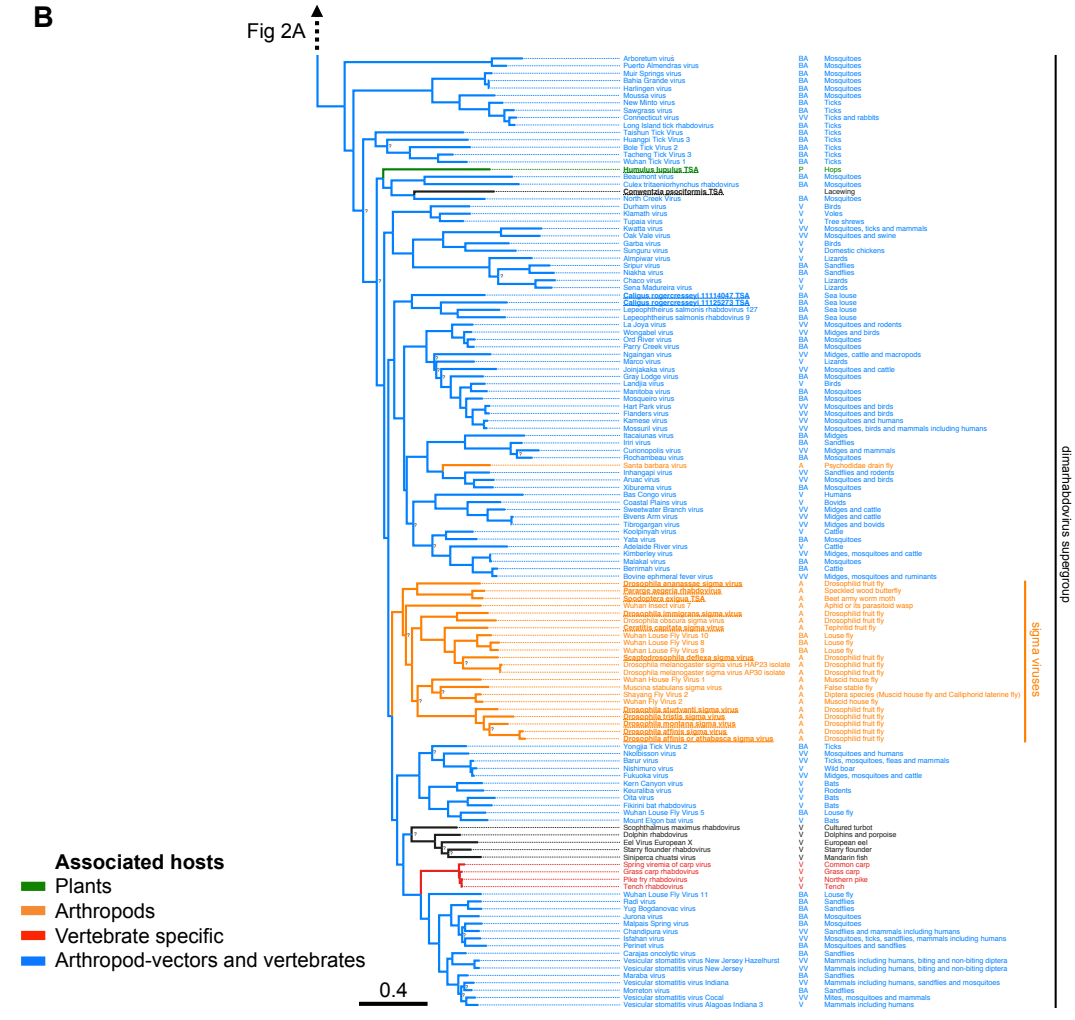
With a few exceptions, rhabdoviruses are either arthropod-vectored viruses of plants or vertebrates, or are vertebrate- or arthropod- specific. In many cases the only information about a virus is the host from which it was isolated. Therefore, *a priori*, it is not clear whether viruses isolated from vertebrates are vertebrate-specific or arthropod-vectored, or whether viruses isolated from biting arthropods (e.g. mosquitoes, sandflies, ticks, midges and sea lice) are arthropod specific or also infect vertebrates. Likewise, it is not clear whether viruses isolated from sap-sucking insects (all Hemiptera: aphids, leafhoppers, scale insect and mealybugs) are arthropod-specific or arthropod-vectored plant viruses.

By combining data on the ambiguous and known host associations with phylogenetic information, we are able to predict both the ancestral and present host associations of viruses (http://dx.doi.org/10.6084/m9.figshare.1425436). To do this we used a Bayesian phylogenetic analysis that simultaneously estimated the phylogeny and host association of our data. In the analysis we defined our host associations either as vertebrate-specific, arthropod-specific, arthropod-vectored vertebrate and arthropod-vectored plant, or as ambiguous between two of these four states (see Methods).

This approach identified a large number of viruses that are likely to be new arthropod-vectored vertebrate viruses (Figure 2b). 80 of 89 viruses with ambiguous host associations were assigned a host association with strong posterior support (>0.95). Of the 52 viruses found in biting arthropods, 45 were predicted to be arthropod-vectored vertebrate viruses, and 6 to be arthropod-specific. Of the 30 viruses found in vertebrates, 22 were predicted to be arthropod-vectored vertebrate viruses, and 2 were predicted to be vertebrate-specific (both fish viruses). Of the 7 viruses found in plant-sap-feeding arthropods (Figure 2a), 3 were predicted to be plant-associated and 2 arthropod-associated.



422



423

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

(A) shows the basal fish-infecting novirhabdoviruses, an unassigned group of arthropod associated viruses, the plant infecting cyto- and nucleo- rhabdoviruses, as well as the vertebrate specific lyssaviruses. **(B)** shows the dimarhabdovirus supergroup, which is predominantly composed of arthropod-vectored vertebrate viruses, along with the arthropod-specific sigma virus clade. Branches are coloured based on the Bayesian host association reconstruction analysis. Black represents taxa omitted from host-state reconstruction or associations with <0.95 support. The tree was inferred from L gene sequences using the Gblocks alignment. The columns of text are the virus name, the host category used for reconstructions, and known hosts (from left to right). Codes for the host categories are: vs= vertebrate-specific, vv= arthropod-vectored vertebrate, a= arthropod specific, ba = biting-arthropod (ambiguous state), v = vertebrate (ambiguous state) and ap=plant-sap-feeding-arthropod (ambiguous state). Names in bold and underlined are viruses discovered in this study. The tree is rooted with the Chuvirus clade (root collapsed) as identified as an outgroup in [13] but we note this gives the same result as midpoint and the molecular clock rooting. Nodes labelled with question marks represent nodes with aLRT (approximate likelihood ratio test) statistical support values less than 0.75. Scale bar shows number of amino-acid substitutions per site.

To test the ability of these reconstructions to correctly predict host associations we randomly selected 10 viruses with known host associations, and assigned them to an ambiguous state. This analysis correctly returned the true host association for all 10 of the randomly chosen viruses with strong posterior support (mean support 0.97, range 0.92-1.00).

Ancestral host associations and host-switches

Viral sequences from arthropods, vertebrates and plants form distinct clusters in the phylogeny (Figure 2). To quantify this genetic structure we calculated a variant of the F_{st} statistic between the sequences of viruses from different groups of hosts. There is strong evidence of genetic differentiation between the sequences from different host groups (P<0.001, Figure S1 http://dx.doi.org/10.6084/m9.figshare.1495351).

Our Bayesian analysis allowed us to infer the ancestral host association of 178 of 188 of the internal nodes on the phylogenetic tree (support >0.95). A striking pattern that emerged is that switches between major groups of hosts have occurred rarely during the evolution of the rhabdoviruses (Figure 2). There are a few rare transitions on terminal branches (Santa Barbara virus and the virus identified from the plant *Humulus lupulus*) but these could represent errors in the host assignment (e.g. cross-species contamination) as well as recent host shifts. Our analysis allows us to estimate the number of times the viruses have switched between major host groups while accounting for uncertainty about ancestral states, and the tree topology and root. There were only two types of host-switch that we found evidence for. Our best estimate of the number of switches from being an arthropod-vectored vertebrate virus to being arthropod specific was three, although one is a jump on a terminal branch may represent contamination (best modal estimate = 3, median = 3.1, CI's= 1.9–6.1). Similarly, we estimate there have been three transitions from being an arthropod-vectored vertebrate virus to a vertebrate-specific virus (best modal estimate = 3, median = 3.1, CI's= 2.9–5.3). We could not determine the direction of the host shifts into the other host groups.

Vertebrate–specific viruses have arisen once in the lyssaviruses clade [3], as well as at least once in fish dimarhabdoviruses (in one of the fish-infecting clades it is unclear if it is vertebrate-specific or vector-borne from our reconstructions). There has also likely been a single transition to being arthropod-vectored vertebrate viruses in the dimarhabodovirus clade.

Insect-vectored plant viruses have arisen once in the cyto- and nucleo- rhabdoviruses, although the ancestral state of these viruses is uncertain. A single virus identified from the hop plant *Humulus lupulus* appears to fall within the dimarhabdovirus clade. However, this may be because the plant was contaminated with insect matter, as the same RNA-seq dataset contains *COI* sequences with high similarity to thrips.

There are two large clades of arthropod-specific viruses. The first is a sister group to the large plant virus clade. This novel group of largely insect-associated viruses are associated with a broad range of insects, including flies, butterflies, moths, ants, thrips, bedbugs, fleas, mosquitoes, water striders and leafhoppers. The mode of transmission and biology of these viruses is yet to be examined. The second clade of insect-associated viruses is the sigma virus clade [11, 12, 19, 45]. These are derived from vector-borne dimarhabdoviruses that have lost their vertebrate host and become vertically transmitted viruses of insects [10]. They are common in Drosophilidae, and our results suggest that they may be widespread throughout the Diptera, with occurrences in the Tephritid fruit fly *Ceratitis capitata*, the stable fly *Muscina stabulans,* several divergent viruses in the housefly *Musca domestica* and louse flies removed from the skin of bats. For the first time we have found sigma-like viruses outside of the Diptera, with two Lepidoptera associated viruses and a virus from an aphid/parasitoid wasp. All of the sigma viruses characterised to date have been vertically transmitted [10], but some of the recently described viruses may be transmitted horizontally – it has been speculated that the viruses from louse flies may infect bats [49] and Shayang Fly Virus 2 has been reported in two fly species [13] (although contamination could also explain this result).Drosophila sigma virus genomes are characterised by an additional X gene between the P and M genes [45]. Interestingly the two louse fly viruses with complete genomes, Wuhan insect virus 7 from an aphid/parasitoid and *Pararge aegeria* rhabdovirus do not have a putative X gene. The first sigma virus was discovered in *Drosophila melanogaster* in 1937 [50], in the last few years related sigma viruses have been found in other *Drosophila* species and a Muscid fly [10-12, 45]. Here we have found sigma-like viruses in a diverse array of Diptera species, as well as other insect orders. Overall, our results suggest sigma-like viruses may be associated with a wide array of insect species.

Within the arthropod-associated viruses (the most sampled host group) it is common to find closely related viruses in closely related hosts (Figure 1). This is reflected in a positive correlation between the evolutionary distance between the viruses and the evolutionary distance between their arthropod hosts (Pearson's correlation=0.36, 95% CI's=0.34-0.38, P<0.001, Figure 3 and Figure S2 http://dx.doi.org/10.6084/m9.figshare.1495351). Since the virus phylogeny is incongruent with that of the respective hosts, this suggests rhabdoviruses preferentially host shift between closely related species [51, 52].

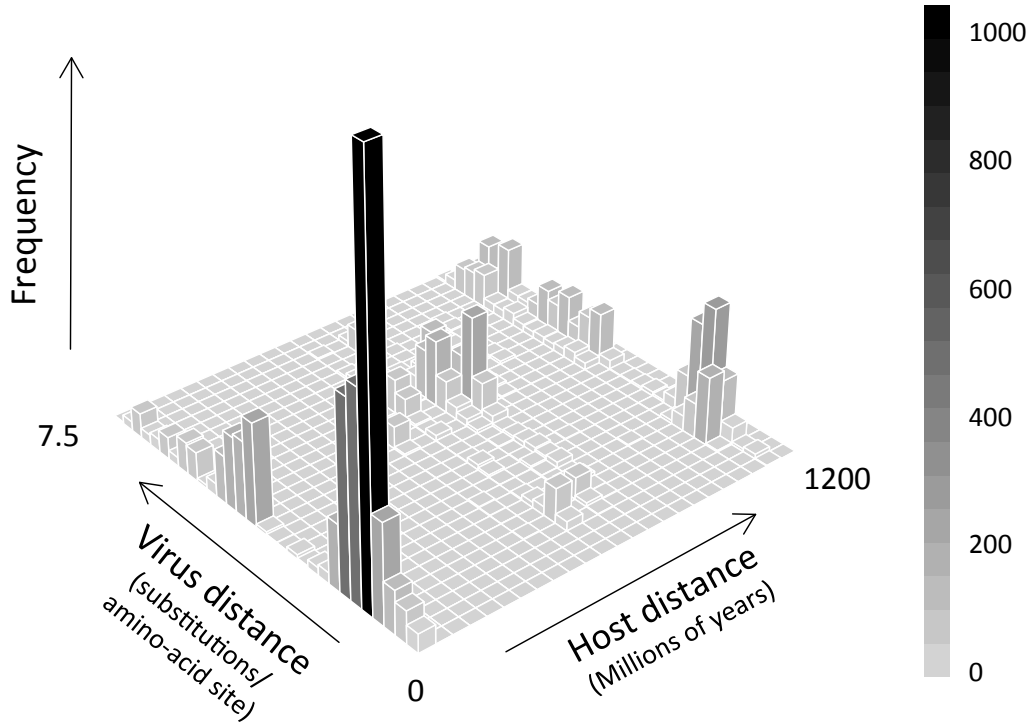


Figure 3. The relationship between the evolutionary distance between viruses and the evolutionary distance between their arthropod hosts (categorised by genus). Closely related viruses tend to be found in closely related hosts. Permutation tests find a significant positive correlation (correlation=0.36, 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

identified a large number of new likely vector-borne viruses – of 85 rhabdoviruses identified from vertebrates or biting insects we predict that 76 are arthropod-borne viruses of vertebrates (arboviruses). By assigning viruses with known associations to an ambiguous state, we have demonstrated that this analysis is robust and can correctly predict the host associations of viruses. Along with the known arboviruses, this suggests the majority of known rhabdoviruses are arboviruses, and all of these fall in a single clade known as the dimarhabdoviruses. In addition to the arboviruses, we also identified two clades of likely insect-specific viruses associated with a wide range of species, suggesting rhabdoviruses may be associated with a wide array of arthropods

We found that shifts between distantly related hosts are rare in the rhabdoviruses, which is perhaps unsurprising as both rhabdoviruses of vertebrates (rabies virus in bats) and invertebrates (sigma viruses in Drosophildae) show a declining ability to infect hosts more distantly related to their natural host [52-54]. It is thought that sigma viruses may sometimes jump into distantly related but highly susceptible species [51, 52, 55], but our results suggest that this rarely happens between major groups such as vertebrates and arthropods. It is nonetheless surprising that arthropod-specific viruses have arisen rarely, as one might naively assume that there would be fewer constraints on vector-borne viruses losing one of their hosts. Within the major clades, closely related viruses often infect closely related hosts (Figure 2). For example, within the dimarhabdoviruses viruses identified from mosquitoes, ticks, *Drosophila*, Muscid flies, Lepidoptera and sea-lice all tend to cluster together (Figure 2B). However, it is also clear that the virus phylogeny does not mirror the host phylogeny, and our data on the clustering of hosts across the virus phylogeny suggest that following major transitions between distantly related host taxa, viruses preferentially shift between more closely related species (Figures 3, S1 and S2) in the same environment (Figure S3).

There has been a near four-fold increase in the number of recorded rhabdovirus sequences in the last five years. In part this may be due to the falling cost of sequencing transcriptomes [56], and initiatives to sequence large numbers of insect and other arthropods [35]. The use of high-throughput sequencing technologies should reduce the likelihood of sampling biases associated with PCR based discovery where people look for similar viruses in related hosts. This suggests that the pattern of viruses forming clades based on the host taxa they infect is likely to be robust. However, these efforts are disproportionately targeted at arthropods, and it is possible that there may be a great undiscovered diversity of viruses in other organisms [57].

Rhabdoviruses infect a diverse range of host species, including a large number of arthropods. Our search has unearthed a large number of novel rhabdovirus genomes, suggesting that we are only just beginning to uncover the diversity of these viruses. The host use by these viruses has been highly conserved during their evolution, which provides a powerful tool to identify previously unknown arboviruses. Given the large number of viruses currently being discovered through metagenomic studies [13, 58, 59], in the future we will be faced by an increasingly large number of viral sequences with little knowledge of the biology of the virus. Our phylogenetic approach can be extended to predict key biological traits in other groups of novel pathogens where our knowledge 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).

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

Figures S1-S3: http://dx.doi.org/10.6084/m9.figshare.1495351
L gene sequences fasta: http://dx.doi.org/10.6084/m9.figshare.1425067
TrimAl alignment fasta: http://dx.doi.org/10.6084/m9.figshare.1425069
Gblocks alignment fasta: http://dx.doi.org/10.6084/m9.figshare.1425068
Phylogenetic tree Gblocks alignment: http://dx.doi.org/10.6084/m9.figshare.1425083
Phylogenetic tree TrimAl alignment: http://dx.doi.org/10.6084/m9.figshare.1425082
BEAST alignment fasta: http://dx.doi.org/10.6084/m9.figshare.1425431
BEAUti xml file: http://dx.doi.org/10.6084/m9.figshare.1431922
Bayesian analysis tree: http://dx.doi.org/10.6084/m9.figshare.1425436
Tables S1-5. List of newly discovered viruses:
http://dx.doi.org/10.6084/m9.figshare.1502665