Molecular Phylogenetics and Evolution 65 (2012) 251–258

Contents lists available at SciVerse ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Phylogeny, integration and expression of sigma virus-like genes in *Drosophila*

Matthew J. Ballinger *, Jeremy A. Bruenn, Derek J. Taylor

Department of Biological Sciences, The State University of New York at Buffalo, Buffalo, NY 14260, USA

ARTICLE INFO

Article history:
Received 24 April 2012
Revised 7 June 2012
Accepted 14 June 2012
Available online 26 June 2012

Keywords:
NIRV
Rhabdovirus
Virus–host interaction
Retrotransposons
Genome evolution
Paleovirology

ABSTRACT

The recent and surprising discovery of widespread NIRVs (non-retroviral integrated RNA viruses) has highlighted the importance of genomic interactions between non-retroviral RNA viruses and their eukaryotic hosts. Among the viruses with integrated representatives are the rhabdoviruses, a family of negative sense single-stranded RNA viruses. We identify sigma virus-like NIRVs of *Drosophila* spp. that represent unique cases where NIRVs are closely related to exogenous RNA viruses in a model host organism. We have used a combination of bioinformatics and laboratory methods to explore the evolution and expression of sigma virus-like NIRVs in *Drosophila*. Recent integrations in *Drosophila* provide a promising experimental system to study functionality of NIRVs. Moreover, the genomic architecture of recent NIRVs provides an unusual evolutionary window on the integration mechanism. For example, we found that a sigma virus-like polymerase associated protein (P) gene appears to have been integrated by template switching of the blastopia-like LTR retrotransposon. The sigma virus P-like NIRV is present in multiple retroelement fused open reading frames on the X and 3R chromosomes of *Drosophila yakuba* – the X-linked copy is transcribed to produce an RNA product in adult flies. We present the first account of sigma virus-like NIRVs and the first example of NIRV expression in a model animal system, and therefore provide a platform for further study of the possible functions of NIRVs in animal hosts.

1. Introduction

Non-retroviral integrated RNA viruses (NIRVs) are a recently discovered form of paleovirus found in eukaryotic genomes (Crochu et al., 2004; Frank and Wolfe, 2009; Horie et al., 2010; Liu et al., 2010; Taylor and Bruenn, 2009; Taylor et al., 2010). NIRVs differ from other paleoviruses, such as endogenous retroviruses, in their need to co-opt reverse transcription machinery. Although retroelements are suspected as the source of reverse transcription machinery, the antiquity of known NIRVs has precluded detailed insights into the evolutionary mechanism of NIRV formation. Still, NIRVs can potentially provide detailed insights about the timescale of host–virus interactions, host–virus coevolution, cryptic host species, and RNA virus diversity. RNA viruses that establish long-term or persistent infections of the host, such as totiviruses and mononegaviruses seem especially well represented in the NIRV paleoviral record (Taylor and Bruenn, 2009; Taylor et al., 2010, 2011). Although most NIRVs are pseudogenes, some have complete open reading frames (ORFs) (Taylor et al., 2011) and are expressed as RNA (Taylor and Bruenn, 2009). Antiviral functions have been proposed for some NIRV groups, but this function has yet to be directly tested.

A roadblock to functional studies of NIRVs is the lack of a multicellular experimental system. Also, there appears to be few cases where a NIRV is closely related to a known RNA virus, complicating the design of effective probes for NIRVs. As part of a large-scale BLAST search of genomic databases, we identified possible sigma virus (Rhabdoviridae) NIRVs within several *Drosophila* species. Katzourakis and Gifford (2010) reported rhabdovirus-like NIRVs using BLAST methods from the *Ixodes* ticks, and *Culex* and *Aedes* mosquitoes. Fort et al. (2011) have expanded the diversity of known hosts and they present sequence evidence to suggest host co-option of the viral polymerase fragment in *A. aegypti*. These authors expected to find sigma virus (SIGMAV)-like NIRVs in *Drosophila* because this virus maintains a persistent infection in host populations by transmission through the germ line (Brun and Plus, 1980), but they found no evidence of integration. Here, we identify and isolate SIGMAV-like NIRVs in this model animal system and test for their expression.

Rhabdoviruses are non-segmented single-stranded negative sense RNA viruses belonging to the order Mononegavirales. Genome size ranges from 11 to 15 kbp, and there are five standard genes present in all rhabdovirus genomes (3′ - N P M G L - 5′), while the presence or absence of additional accessory genes varies between genera (Walker et al., 2011). Rhabdoviruses are commonly vectored by arthropods; hosts include a wide variety of plants, invertebrates, and vertebrates (Kuzmin et al., 2009). Major genera of rhabdoviruses are *Novirhabdovirus, Vesiculovirus, Ephemerovirus,*

* Corresponding author.
E-mail addresses: ballinge@buffalo.edu (M.J. Ballinger), cambruen@buffalo.edu (J.A. Bruenn), djtaylor@buffalo.edu (D.J. Taylor).

1055-7903/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ympev.2012.06.008

Cytorhabdovirus, Nucleorhabdovirus, and *Lyssavirus* (Tordo et al., 2005). *Vesiculoviruses* and *Ephemeroviruses* together form the Dimarhabdoviruses (Bourhy et al., 2005). Rhabdoviruses are well known as pathogens (e.g. rabies in mammals), and have at least two well-developed experimental systems in SIGMAV and vesicular stomatitis virus. However, the evolutionary context and diversity of many rhabdovirus groups is poorly studied and would likely benefit from paleoviral information. Recently, phylogenetic analyses have shown that the *Drosophila* SIGMAVs, including *Drosophila melanogaster* SIGMAV (DMelSV), *Drosophila obscura* SIGMAV (DObsSV), *Drosophila affinis* SIGMAV (DAffSV), *Drosophila tristis* SIGMAV (DTrisSV), *Drosophila immigrans* SIGMAV (DImmSV), *Drosphila ananassae* SIGMAV (DAnaSV), and one SIGMAV infecting a non-Drosophilid, *Muscina stabulans* SIGMAV (MStaSV), form a sister clade to the Dimarhabdoviruses (Longdon et al., 2010, 2011b). The SIGMAVs are vertically transmitted and seem to have recently swept through natural populations of *Drosophila melanogaster* (Brun and Plus, 1980; Carpenter et al., 2007; Longdon et al., 2011a). Infection is easily phenotyped in the laboratory, as infected flies die or experience permanent paralysis upon CO_2 exposure (L'Heritier, 1958).

In this report, we further expand the existing evidence of rhabdovirus NIRVs in arthropods and show that evolutionary interactions between the *Drosophila* SIGMAVs and *Drosophila* are more widespread than previously known. We isolate NIRVs from *Drosophila spp*., identifying a potential multicellular experimental model for studying the biology of NIRVs. Furthermore, we describe a case of recent NIRV integration that exhibits compelling evidence for a retrotransposon-mediated mechanism. Finally, we present evidence for RNA-level expression of a NIRV in *Drosophila yakuba*.

2. Materials and methods

2.1. Nucleic acid extraction and PCR

Live cultures of *Drosophila willistoni* (14030-0811.24) and *D. yakuba* (14021-0261.01) were obtained from the *Drosophila* Species Stock Center at University of California, San Diego and DNA was extracted from each using the Qiagen DNeasy Blood & Tissue Kit to manufacturer's specifications. For PCR, primers were designed targeting the flanking sequences of each NIRV. The *D. willistoni* N-like NIRV (PCR targeted region 1.7 kb) was amplified using 2 primer sets, WIL1 (5′-CTATTGCACTATCGGGAGTGTGGC, 5′-GTCCAAAACTGACATTAACATCGGC, 830 bp product) and WIL2 (5′-GGGTTCAACTTCATCAACATCGGC, 5′-GCACATGTTAATGTCAGTTTTGGAC, 948 bp product). WIL1 and WIL2 reactions were run for 40 cycles of 94 °C denaturation, 57 °C annealing, and 72 °C extension. The *D. yakuba* L-like NIRV (PCR targeted region 785 bp) was amplified using the primers 5′-AATACATCTGCCTGCTGTCTTGGC and 5′-GACTAAGATTT GTGTTTCCCGTGC, and the reaction was run for 35 cycles of 94° denaturation, 55 °C annealing, and 72 °C extension. DNA template used for RT-PCR was extracted using the Qiagen DNeasy Blood & Tissue Kit to manufacturer's specifications. RNA template used for RT-PCR was extracted with BioBasic's total RNA extraction kit and treated with an extra step of DNase I incubation (Promega) for 30 min at 37 °C. The P-like NIRV primers used for RT-PCR were 5′-GCTCTACTATGGACTCGGAATCAG and 5′- ATCTAAGCATCATAGTGGAGGGAGC (765 bp product). Actin primers for positive control are 5′-ATGTGTGACGAAGAAGTTGCTGC and 5′-GTGTTGGCATACA GATCCTTACG (890 bp product). RT-PCR on the *D. yakuba* P-like NIRV was done using Qiagen's Onestep RT-PCR kit and thermal cycling consisted of an initial reverse transcription step at 50 °C for 30 min, a Taq activation step at 94 °C for 15 min, followed by 35 cycles of 94 °C denaturation for 30 s, 50 °C annealing for 30 s, and 72 °C extension for 1 min 30 s. PCR and RT-PCR products were gel-purified and sequenced by the DNA Sequencing Facility at Roswell Park Cancer Institute. The sequence of the *D. yakuba* P-like NIRV transcript was submitted to GenBank under GenBank ID: JN093015.

2.2. Bioinformatics

tBLASTn searches of NCBI's *Drosophila* whole-genome shotgun databases (http://www.ncbi.nlm.nih.gov/) were performed using the query sequences given in Table S1. Translated nucleotide sequences from matches of high significance (expected value $<10^{-5}$) were retained for alignment. We retained sequences greater than 100 amino acids to minimize complications in phylogenetic tests associated with short sequences. To identify blastopia elements in *Drosophila*, we used the blastopia polyprotein sequence published by Frommer et al. (GenBank ID: CAA81643) as a query in a tBLASTn search of the *Drosophila* WGS databases.

MAFFT (Katoh et al., 2002) was used to create alignments using the default parameters for amino acids. Ambiguous regions of the alignments were filtered out with GBlocks (Castresana, 2000) or Guidance (Penn et al., 2010). Maximum-likelihood trees were built with RAxML (Stamatakis et al., 2008) on the CIPRES Science Gateway (Miller et al., 2010) and PhyML (Guindon et al., 2010) using the most appropriate models of amino acid substitution for the gene trees as determined by ProtTest (Abascal et al., 2005). These were LG + I + G + F for the L sequences, LG + I + G for the N sequences, and JTT + G for the P sequences. We used the Bayesian information criterion (BIC) for model selection with ProtTest, as it has been shown to outperform the Akaike information criterion (AIC) on simulated data sets (Zhang et al., 2010). RAxML was used to estimate the number of bootstrapping pseudoreplicates for all trees.

3. Results and discussion

We identified sigma- and other rhabdovirus-like NIRVs in *Drosophila* by tBLASTn searches of whole-genome shotgun (wgs) databases, using rhabdovirus RNA-dependent RNA polymerase (L), nucleocapsid protein (N), and polymerase-associated protein (P) amino acid sequences as queries (Table S1). As described in a previous report (Katzourakis and Gifford, 2010), rhabdovirus-like NIRVs are present in *Ixodes scapularis, Aedes aegypti*, and *Culex quinquefasciatus*, as well as salmon lice, sand flies, flour beetles, and one *Drosophila* species (*D. sechellia*) (Fort et al., 2011). The latter authors note that the SIGMAVs of *Drosophila* represent a promising candidate for NIRV formation because the virus is maintained in natural host populations by transmission through the germ line (Brun, 1977), but their searches failed to identify SIGMAV-like NIRVs. Our searches revealed SIGMAV-like and other rhabdovirus-like NIRVs in *Drosophila willistoni, D. yakuba, D. virilis, D. grimshawi, D. eugracilis*, *D. sechellia, D. rhopaloa, and D. biarmpes*. As expected, many of these NIRVs contain ORF disruptions, consistent with direction of transfer from virus to animal.

To address concerns that these matches might be assembly artifacts, we examined NCBI Genomic Trace Archives for these regions. In all cases for which traces are available, NIRV loci were supported with multiple and overlapping coverage of at least 95% sequence identity. To firmly establish NIRV presence, we extracted DNA and performed PCR on each NIRV type in *Drosophila*. N-like amplicons were isolated from *D. willistoni* genomic template, and L- and P-like amplicons from *D. yakuba*.

3.1. Phylogenetic evidence for a SIGMAV origin

Phylogenetic analyses of L-like and N-like (Figs. 1 and 2) sequence alignments confirmed that our BLAST sequences are closely related to SIGMAV. The L phylogeny supports SIGMAV L-like NIRVs in *D. yakuba* and *D. eugracilis*. Both *D. yakuba* and *D. eugracilis*

belong to the melanogaster subgroup, but their infecting viruses do not mirror this relationship, supporting the conclusions of Longdon et al. (2011b) regarding host switching by SIGMAV. A third NIRV in *D. virilis* more likely originated during infection by a plant rhabdovirus related to the Nucleorhabdoviruses. The L-like NIRVs in each of these species contain ORF disruptions.

The N phylogeny supports DMelSV-like NIRVs in *D. willistoni* and *D. grimshawi*, while two more, one in D. *sechellia* and a second in D. *willistoni* are more closely related to the N sequences of DObsSV. There are also N-like NIRVs in *D. rhopaloa* and *D. biarmpes* which may have not originated from SIGMAV infections at all; rather these sequences are more closely related to members of the Vesiculovirus clade. It should be noted, however, that only two SIGMAV N sequences are known; an underrepresentation of the existing diversity infecting natural *Drosophila* populations. With the exception of *D. biarmpes*, the N-like NIRVs in these species of *Drosophila* are present in at least one copy containing a preserved ORF. We investigated the expression of these putative coding regions bioinformatically by blasting EST, RNA-Seq, and non-redundant (nr) protein databases for each species (for *D. rhopaloa*, the RNA-Seq database is available at the HGSC modENCODE genome project website, also see Table S1). Only *D. rhopaloa* returned a corresponding RNA copy of the N-like NIRV, and it is present in the RNA-Seq database produced from eggs, but not in databases from either sex of the adult fly.

Drosophila yakuba was the only species to return matches to a SIGMAV P protein amino acid query. We identified three copies of this NIRV mapped to chromosomes in *D. yakuba*, and two more on short, unmapped contigs. Of the mapped copies, two are present on chromosome 3R and one on the X chromosome. Phylogenetic analysis places this NIRV with the SIGMAV group (Fig. 3). In comparison to the other NIRVs in *Drosophila* spp., the *Drosophila yakuba* SIGMAV P-like NIRV shows elevated sequence similarity with the putative viral source, implying that it is either a more recent integration or the sites have been selectively conserved. Notably, a BLAST of the *D. yakuba* RNA reference sequence database revealed an unnamed RNA product that is highly similar to this NIRV (FlyBase ID: FBgn0232275). For these reasons, we selected the P-like NIRV as a target for reverse transcriptase-PCR (RT-PCR). As shown in Fig. 4, RT-PCR on RNA extracted from adult *D. yakuba* flies confirmed that this NIRV is transcribed. The nucleotide sequences of the 3R and X-linked NIRVs align with that of DMelSV P gene coding region, but adequate diversity of SIGMAV sequences for further study of the evolutionary history of this NIRV is currently lacking. While many DMelSV genomic strains have been sequenced, very little genetic differentiation exists between these, likely due to the recent sweep (Carpenter et al., 2007). The only other known SIGMAV P sequence comes from DObsSV and is highly divergent.

3.2. Sequence evidence for integration by retrotransposon template-switching

Retrotransposon activity has been demonstrated in an experimental setting to be an effective integration mechanism for exogenous non-retroviral RNA (Geuking et al., 2009). Where possible, we have examined the neighboring sequences of SIGMAV-like NIRVs. In *D. sechellia*, *D. yakuba*, and *D. willistoni*, non-LTR retroelement coding regions are found adjacent to or flanking the NIRV sequence (Table S2). This association is far from ubiquitous in our data set, and for the contigs in which retroelements were found, their sequences and those of the NIRVs are eroded and fragmented to such a degree that we are unable to argue decisively as to their integration mechanism. Indeed, the prevalence of mobile elements in animal genomes leaves room for this handful of associations to be

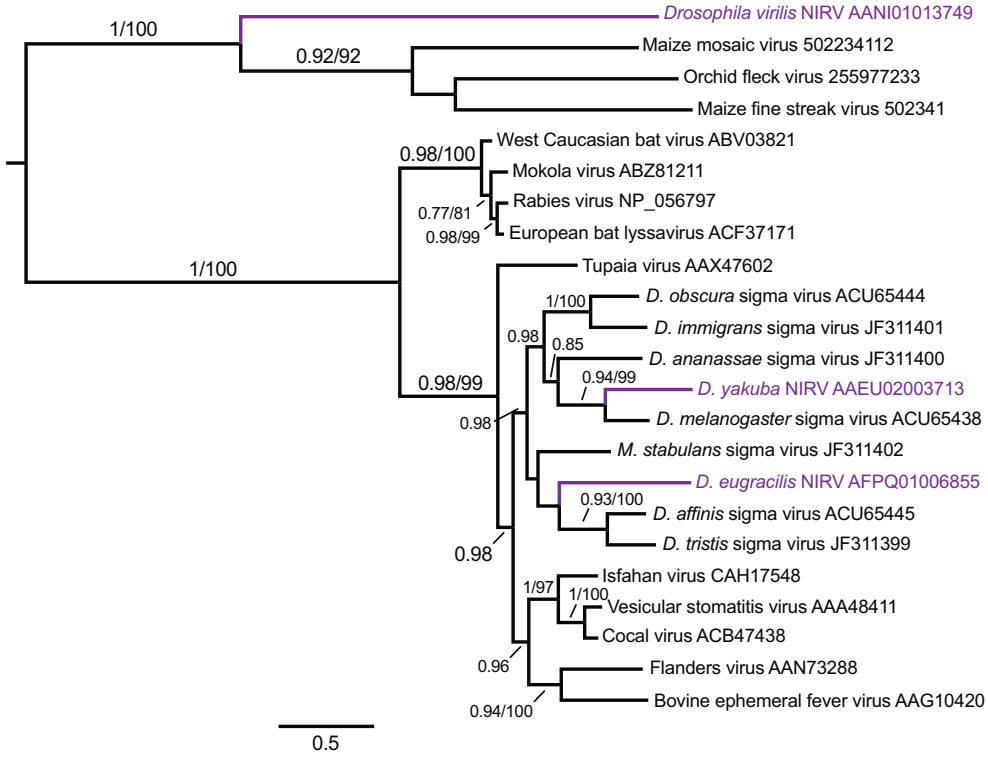

Fig. 1. Midpoint rooted maximum likelihood phylogram of RNA-dependent RNA ploymerase (L) amino acid sequences from rhabdoviruses and related genomic sequences of *Drosophila*. Branch support values greater than .75 for likelihood ratio tests and 75 for bootstrapping are listed in respective order on branches. Branches and annotations for L-like related sequences from *Drosophila* **spp.** are shown in purple. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

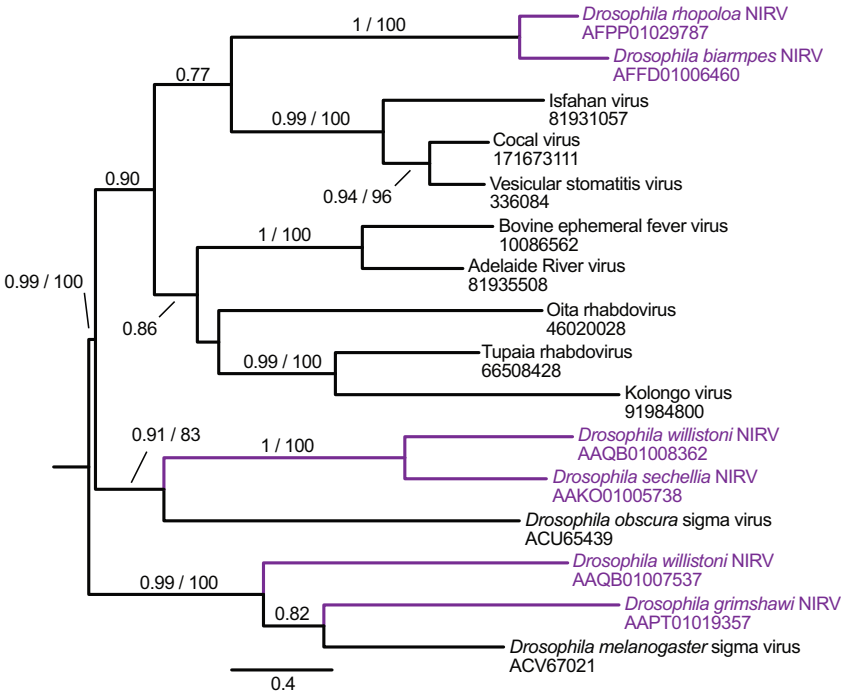

Fig. 2. Midpoint rooted maximum likelihood phylogram of nucleocapsid protein (N) amino acid sequences from rhabdoviruses and related genomic sequences of *Drosophila*. Branch support values greater than .75 for likelihood ratio tests and 75 for bootstrapping are listed in respective order on branches. Annotations for N-like related sequences from *Drosophila* **spp.** are shown in purple. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

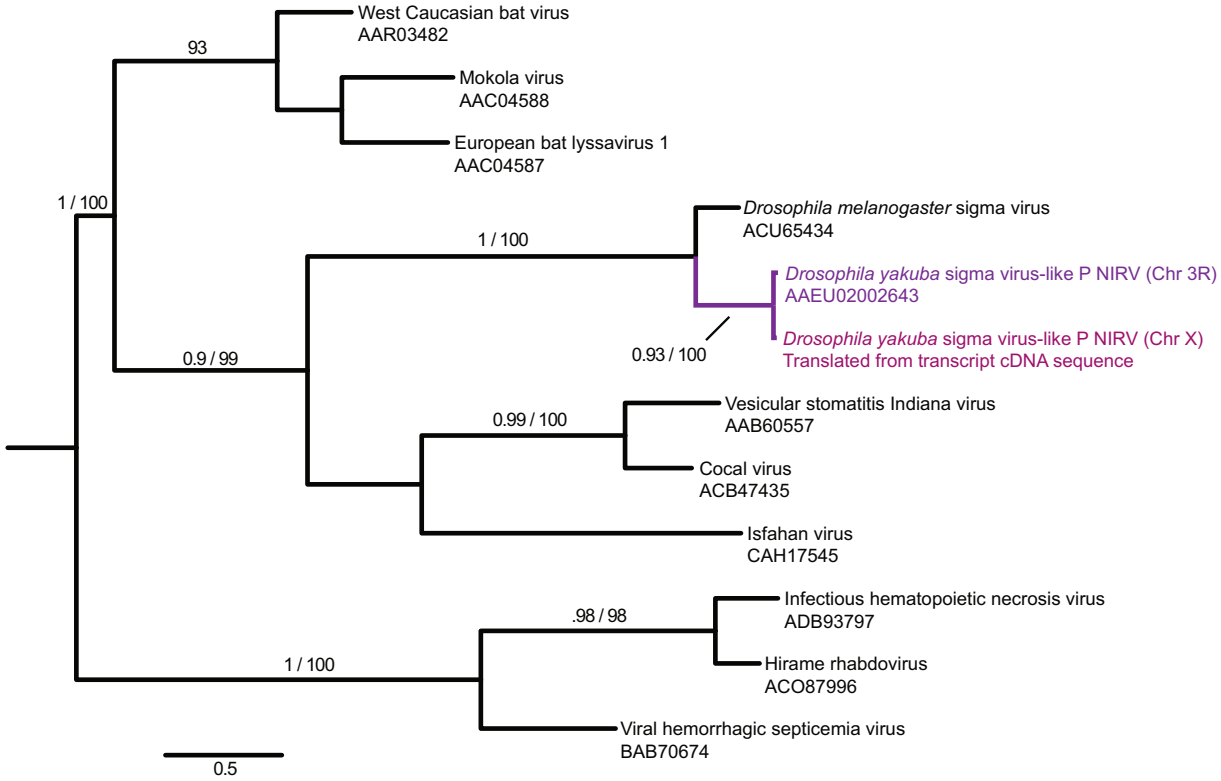

Fig. 3. Midpoint rooted maximum likelihood phylogram of polymerase-associated protein (P) amino acid sequences from rhabdoviruses and translated genomic and cDNA sequences from *Drosophila yakuba*. Branch support values greater than .75 for likelihood ratio tests and 75 for bootstrapping are listed in respective order on branches. Branches and annotations for the P-like NIRVs from *Drosophila yakuba* are shown in purple. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

interpreted as coincidental. However, we are able to present compelling evidence for retrotransposon-mediated integration in the case of the SIGMAV P-like NIRVs in *D. yakuba*. By mapping the immediate neighboring regions of chromosome 3R and X, we have identified sequences similar to the coat and nucleoprotein coding regions of a blastopia LTR retrotransposon juxtaposed to and in

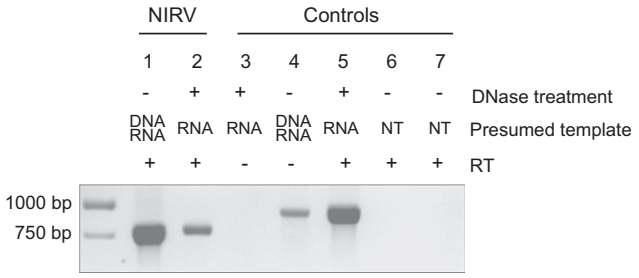

Fig. 4. DNA gel electrophoresis of reverse transcription PCR (RT-PCR) products confirms an expression product for the *Drosophila yakuba* P-like NIRV. RT-PCR products were loaded into a 1.5% agarose in .5× TBE gel to confirm the presence of a P-like NIRV-derived RNA transcript. A 1 kb ladder (Fermentas) was used as a standard (shown on the left side of the gel image). Primers target 765 bp of the P-like NIRV open reading frame (ORF). Reaction-specific treatments are designated by + or − and are explained at right. Taq polymerase was present in all reactions. Reactions in lanes 1, 2, 3, and 6 were amplified with NIRV specific primers designed to amplify from each of the three known P-like NIRVs in *D. yakuba*. Reactions in lanes 4, 5, and 7 were amplified with actin-specific primers. All lanes in which the presumed template is designated as RNA used a nucleic acid extraction protocol specific to RNA, and were additionally treated with DNase I. Lane 3 is a negative control for DNA contamination. Lane 4 is a positive control for the reagents in lane 3. NT, no template.

reading frame with the SIGMAV P-like NIRV, creating a blastopia gag-SIGMAV P gene fusion. We then used the LTR_finder web server (Xu and Wang, 2007) to locate the flanking LTRs, and BLASTs to identify additional sequences with high similarity to blastopia polyprotein (Fig. 5). Blastopia elements belong to the Ty3/Gypsy class of LTR retrotransposons (Malik and Eickbush, 1999) and are active in *Drosophila* under the regulatory control of bicoid during early development (Frommer et al., 1994). We performed BLAST searches to confirm these elements are present in the *D. yakuba* genome at high copy number. The structure of this fused NIRV suggests template switches during reverse transcription of the retroelement, resulting in a single ORF encoding a novel putative protein. An additional, partial blastopia element is present in tandem copy on chromosome 3R and contains the third copy of the P-like NIRV, but the upstream blastopia coding regions and LTR are missing from this retroelement fragment. The target site duplication (TSD) of the X-linked NIRV-containing element differs from that of the 3R element, suggesting that the ancestral mobile element maintained some degree of transcriptional and integrative viability following the integration of the original NIRV-containing element.

The addition of entire ORFs to existing retroelements, while maintaining and even expanding element function has been described, for example, in the case of envelope gene acquisition by endogenous retroelements leading to the emergence of new retroviruses (Malik et al., 2000). Note, however, that the blastopia reverse transcriptase (RT) coding region is absent in the NIRV-containing elements, so another RT must be commandeered in order for these elements to complete the replication cycle. Both full copies of this NIRV-containing element contain a six base pair non-consensus bicoid homeodomain recognition sequence (TGATCC) in the 5′ LTR, and all three copies in the 3′ LTR (Zhao et al., 2000). Further characterization of this putative regulatory region and its role in the expression of the NIRV-containing element is an aim for future studies. In the present study, we used multiple primer sets (see Fig. 5) in our attempts to amplify a transcript for the blastopia-NIRV fused ORF; however, only the primer set that targeted the NIRV portion of this coding region successfully produced a transcript-derived product during RT-PCR.

Fig. 5 maps out the general features of these elements and shows for comparison the structure of a blastopia element found on the *D. yakuba* 3R chromosome that does not contain a SIGMAV P-like NIRV. Of particular note is that the SIGMAV P-like NIRV coding sequence is in frame with the coat and nucleoprotein components of the retroelement gag gene. Most of the pol components are absent entirely, with the exception of integrase (IN), which is present at high amino acid site identity to *D. melanogaster* blastopia integrase (expected value $2e^{-89}$, 47% identical sites). The pseudogenized reverse transcriptase gene present between the 3R element and its partial duplication is a non-LTR RT related to the *Drosophila* LINE, X-element (Tudor et al., 2001).

In order to produce the observed blastopia/NIRV fusion structure, two template switches are required, the first from the LTR element RNA to the sigma virus P RNA, and a second, back to the LTR element. We note that only a complete element containing two LTRs can be selected for integration, therefore the observed structure should be unsurprising. The second switch occurred immediately following the poly-A tail of sigma virus P; at this position, the sequence returns to the blastopia IN coding region. Rhabdoviruses polyadenylate their mRNAs by RdRp stuttering during transcription at a stretch of seven uracils present after each ORF in the genomic template (Li et al., 2009; Tekes et al., 2011). The presence and position of the poly-A tail in this NIRV sequence indicates that it must have been generated from a SIGMAV P mRNA. This is supported by the sequence at the 5′ end of the NIRV. In rhabdovirus genomes, each gene is preceded by a conserved transcription initiation sequence and a 5′ UTR that varies in size. In SIGMAV, the sequence CAACANC precedes the P, X, M, G, and L genes (Contamine and Gaumer, 2008), and in the DMelSV P gene, it is followed by a 47 base pair UTR (GenBank ID: HQ655096). In the NIRV, AAACACC is present at the extreme 5′ end, 47 base pairs upstream of the SIGMAV P start codon (Fig. S1). The NIRV positions corresponding to the 5′ UTR and the first 36 codons of the viral P mRNA are the least conserved in comparison with the viral sequence, though the match is still significant (the tblastx expected value for alignment of these residues is $4e^{-12}$). Presumably, this is attributable to diminished functional constraint on these residues, but we currently have no means to estimate how much of this sequence divergence was present in the *D. yakuba*-infecting SIGMAV versus that which has developed since host integration. Addition of sequences upstream of the SIGMAV initiation site into the tblastx alignment results in exclusion of those sequences and an increased expected value. The NIRV copies show a surprisingly high degree of similarity with the viral RNA, indicating that the *D. yakuba*-infecting SIGMAV is much more closely related to DMelSV than any other known SIGMAV. However, there are some notable differences between the NIRV copies themselves. For instance, the sequence at the site of the first template switch has been lost from the X-linked NIRV-containing element (despite this deletion, the blastopia/SIGMAV P ORF is maintained). This region is preserved in the 3R element with such high similarity to the original contributing sequences that the site of the first template switch can be precisely assigned. The overall effect of the lost sequences is that the 3R element contains a much longer fused ORF. The single case in which the X-linked NIRV has acquired an insertion – 10 bp containing a premature stop codon at the 3′ end – results in truncation of the putative protein's C-terminus by 41 amino acids. These 10 bp are unique to the X-linked copy and the sequenced transcript, indicating that the X-linked copy is the coding region for the expressed NIRV (Fig. S2). Unexpectedly, the transcript sequence we have produced does not match that of the previously mentioned transcript available on FlyBase. From the 5′ end, the first 712 bp of the FlyBase transcript sequence match the 3R NIRV; however, the final 110 bp at the 3′ terminus of this transcript diverge from the SIGMAV derived NIRV sequences. These nucleotides encode 2 CCHC zinc-finger type DNA binding domains that match those encoded by another retroelement (X element-like). We considered that this fused transcript may be encoded by an additional P-like NIRV that we missed in our initial searches, and we performed additional blasts against the *D. yakuba* WGS

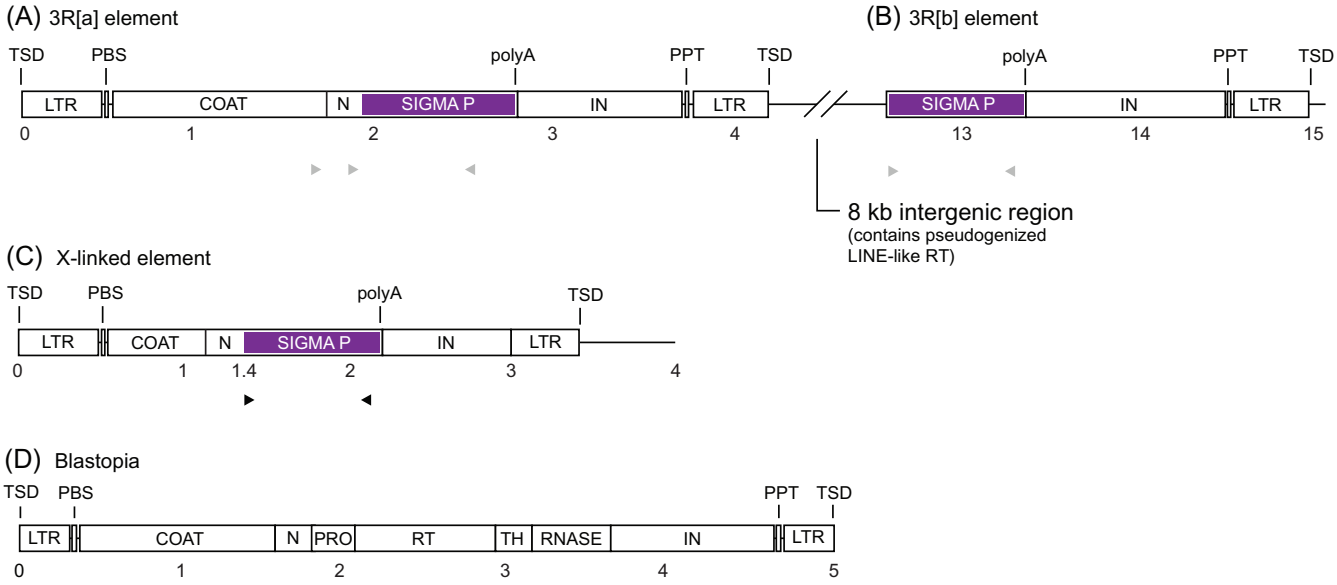

Fig. 5. Genome maps of P-like NIRV-containing blastopia LTR retrotransposons on the 3R and X chromosomes of *Drosophila yakuba*. Number labels below the maps indicate sequence length in kilobases. LTR, long terminal repeat; TSD, target site duplication; PBS, tRNA primer binding sequence; PPT, polypurine tract; RT, reverse transcriptase coding region; TH, tether; IN, integrase coding region; UI, an unidentified ORF of blastopia. Black arrowheads below the maps indicate primer target regions that yielded a transcript product by RT-PCR, while gray arrowheads identify primer targets with no observable RT-PCR products (genomic products were present). (A) The NIRV-containing blastopia element on chromosome 3R. A complete element is present upstream of (B) a partial duplication containing only one LTR and a portion of the protein coding region. (C) The NIRV-containing blastopia element present on the X chromosome. (D) A blastopia element present on chromosome 3R that lacks the integrated NIRV. A multiple sequence alignment of the NIRV-containing elements in this figure is shown in *Fig.* S1.

database in an attempt to map this coding region, but the transcript has no corresponding coding region in the final *D. yakuba* genome assembly. Because the transcript can be wholly attributed to *D. yakuba* genomic sequences, but it cannot be traced back to a single coding region, the simplest explanation is that the relevant contigs are artifacts created during library production, sequencing or assembly that failed to meet length or coverage cutoffs to be included in the final assembly. A problem with this interpretation is that the fused sequence is present in both the WGS and RNA-Seq databases, which were created from independent libraries. As such, the only realistic explanations are: (1) the coding region was excluded from the genome assembly in error, or (2) the genome assembly is correct, and this transcript is a recombinant of two RNAs that was subsequently reverse transcribed *in vivo*. We note that the second scenario requires transcription of the P-like NIRV on chromosome 3R, an event we were unable to support by RT-PCR.

The discovery that the NIRV-containing blastopia elements are present in multiple genomic copies raises questions regarding the direction of inheritance and the age of each element. The replication mechanism used by LTR retrotransposons presents a unique opportunity to address these questions. With the possible exception of mutations incurred during transposition, an LTR retrotransposon is expected to have identical LTR sequences at the time of integration. From that point, these sequences are expected to diverge at the neutral mutation rate such that the divergence between LTRs of a given element might be used to estimate its age. Alignment of LTRs of the NIRV-containing element on the *D. yakuba* X chromosome reveals four nucleotide differences, while the alignment of the LTRs flanking the 3R NIRV-containing element reveals two. With so few mutations accumulated between LTRs both within and between elements, it is difficult to rank ages with confidence. The elevated base misincorporation rate of reverse transcriptase further complicates this approach, possibly accounting for a large portion of divergent nucleotides. As we have no means of differentiating between transposition-induced mutations and those arising post-integration, we are limited to the upper bound in estimating the age of these elements. Incorporating a neutral mutation rate of 5.8×10^{-9} per site per generation (Haag-Liautard et al., 2007), and an average of 7 generations per year for *Drosophila*, yields an approximate age of 118,000 years for the X-linked NIRV-containing element and 59,000 years for the 3R element (there is no known difference in autosome versus X chromosome neutral mutation rates in *Drosophila*) (Vicoso and Charlesworth, 2006). Under a neutral accumulation perspective, the X-linked element is older and therefore the direction of copy must have been out of the X chromosome, consistent with a known general bias of gene movement from X to autosomes in *Drosophila* (Vibranovski et al., 2009b). We acknowledge that values for mutation rate and generation time that reflect those occurring in natural populations are difficult to obtain with confidence, and are subject to fluctuations through evolutionary time. However, in this context, incorporating any range of plausible values for these variables leads to the same conclusion; both SIGMAV P-like NIRV-containing elements are young and minimally divergent, and therefore this NIRV must be a recent integration.

3.3. Potential NIRV function in the Drosophila/SIGMAV system

While at present there is no evidence to support a function for this expressed SIGMAV P-like NIRV in *D. yakuba*, we consider it to be a unique candidate for further investigation into the hypothesis that NIRVs can play a role in virus–host interaction. Since this NIRV is present on and transcribed from the X chromosome, it should be subject to sex-specific expression patterns and the evolutionary trajectory of X chromosome sequences. As cited above, a trend of gene movement from the X to autosomes by retrotransposition has been reported (Vibranovski et al., 2009b). One adaptive explanation for this is compensation for the inability of males to regulate their already hypertranscribed X chromosome, as well as loss of X chromosome gene expression by inactivation during spermatogenesis, also known as meiotic sex chromosome inactivation (MSCI) (Hense et al., 2007; Kemkemer et al., 2011; Vibranovski et al., 2009a). It is tempting to speculate a similar pressure encouraging the copying of the X-linked NIRV-containing blastopia element to 3R; however, it is also possible that the movement to 3R

was non-adaptive, and the lack of expression from 3R appears to be more consistent with this interpretation. This absent expression has many possible explanations. The simplest are point mutations in the LTRs between the X-linked and 3R elements, differences in putative up and downstream regulatory sequences, and physical position on the respective chromosomes. Still another possibility is that expression does occur from the 3R NIRV, though in a temporally restricted pattern.

In *D. melanogaster*, resistance to SIGMAV is attributed in part to interactions between endogenous gene product Ref(2)P and the sigma P protein, which act to suppress viral replication (Carre-Mouka et al., 2007; Guillemain, 1953). Nonsynonymous substitutions in the PB-1 domain that determine a restrictive versus permissive Ref(2)P allele have been identified, and phylogenetic analysis of presently observed haplotypes supports the accumulation of presumably restrictive mutations from ancestrally permissive alleles, possibly as a co-evolutionary response to corresponding SIGMAV mutations (Carre-Mouka et al., 2007). The precise mechanism by which Ref(2)P yields a resistant phenotype remains unclear; however, physical interaction between Ref(2)P and SIGMAV P protein, and shared antigenic structures between Ref(2)P and SIGMAV N protein have been described (Wyers et al., 1993). These Ref(2)P interactions are reflected in what is known about the P protein in the rhabdovirus life cycle. The viral P protein is necessary for transcription and replication, as it recruits the polymerase to the genomic template. It is modularized in a similar fashion to a transcriptional activator, with specific polymerase-binding and template-binding (P binds to N which is in complex with genomic RNA) domains (Emerson and Schubert, 1987). While it may be premature at this stage to form a specific hypothesis about the role of the P-like NIRV in *D. yakuba*–SIGMAV interactions, the high degree of sequence similarity between SIGMAV P and the translated P-like NIRV makes a role involving the regulation of the SIGMAV life cycle via direct interaction with viral proteins appear promising.

4. Conclusion

We describe SIGMAV-like NIRVs in the genomes of multiple *Drosophila* species. We also present a case in which one of these is being expressed at the RNA level, thereby presenting a possible foundation for experimental investigation of NIRV function in animals. The genomic neighborhood of this integration – its containment and transposition within a blastopia LTR retrotransposon – confirms the suspected role of retroelements in NIRV integration in nature and has provided adequate context to estimate that this integration occurred in the recent past. The synthesis of previous decades of SIGMAV research with our discoveries here provides opportunity to gain new insight into the SIGMAV–*Drosophila* interaction, and potentially virus–host interactions in general, as NIRVs have now been identified across a wide range of eukaryotes.

Our findings establish that sigma virus-like NIRVs are present in *Drosophila* species and that these infection scars represent a rich evolutionary history between virus and host. Although the existence of NIRVs in arthropods has been predicted by the expected exposure of the host to viruses (e.g., the persistent infection of *Drosophila* with sigma virus) exposure rates fail to explain the large variation within host genera in the distribution of NIRVs. As NIRV discovery is dependent on the few available virus and host genome sequences, it is likely that SIGMAV- and other rhabdovirus-like NIRVs in nature extend far beyond the scope of this report.

Acknowledgments

We thank the administration team of the Center for Computational Research (University at Buffalo) for set up, monitoring, and use of the U2 cluster. M.J.B. acknowledges a fellowship from the Center for Advanced Molecular Biology and Immunology (CAMBI).

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2012.06.008.

References

Abascal, F., Zardoya, R., Posada, D., 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21, 2104–2105.
Bourhy, H., Cowley, J.A., Larrous, F., Holmes, E.C., Walker, P.J., 2005. Phylogenetic relationships among rhabdoviruses inferred using the L polymerase gene. J. Gen. Virol. 86, 2849–2858.
Brun, G., 1977. Infection of drosophila female germ line cells by sigma virus. Ann Microbiol A128, 119–131.
Brun, G., Plus, N., 1980. The Viruses of Drosophila. Academic Press, New York.
Carpenter, J.A., Obbard, D.J., Maside, X., Jiggins, F.M., 2007. The recent spread of a vertically transmitted virus through populations of *Drosophila melanogaster*. Mol. Ecol. 16, 3947–3954.
Carre-Mouka, A., Gaumer, S., Gay, P., Petitjean, A.M., Coulondre, C., Dru, P., Bras, F., Dezelee, S., Contamine, D., 2007. Control of sigma virus multiplication by the ref(2)P gene of *Drosophila melanogaster*: an in vivo study of the PB1 domain of Ref(2)P. Genetics 176, 409–419.
Castresana, J., 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17, 540–552.
Contamine, D., Gaumer, S., 2008. Sigma rhabdoviruses. In: Mahy, B.W.J., Van Regenmortel, M.H.V. (Eds.), Encyclopedia of Virology. Elsevier, Oxford, pp. 576–581.
Crochu, S., Cook, S., Attoui, H., Charrel, R.N., De Chesse, R., Belhouchet, M., Lemasson, J.J., de Micco, P., de Lamballerie, X., 2004. Sequences of flavivirus-related RNA viruses persist in DNA form integrated in the genome of Aedes spp. mosquitoes. J. Gen. Virol. 85, 1971–1980.
Emerson, S.U., Schubert, M., 1987. Location of the binding domains for the RNA polymerase-L and the ribonucleocapsid template within different halves of the NS phosphoprotein of vesicular stomatitis-virus. Proc. Natl. Acad. Sci. USA 84, 5655–5659.
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.
Frank, A.C., Wolfe, K.H., 2009. Evolutionary capture of viral and plasmid DNA by yeast nuclear chromosomes. Eukaryot. Cell 8, 1521–1531.
Frommer, G., Schuh, R., Jackle, H., 1994. Localized expression of a novel micropia-like element in the blastoderm of Drosophila melanogaster is dependent on the anterior morphogen bicoid. Chromosoma 103, 82–89.
Geuking, M.B., Weber, J., Dewannieux, M., Gorelik, E., Heidmann, T., Hengartner, H., Zinkernagel, R.M., Hangartner, L., 2009. Recombination of retrotransposon and exogenous RNA virus results in nonretroviral cDNA integration. Science 323, 393–396.
Guillemain, A., 1953. Discovery and localization of a gene in *Drosophila melanogaster* inhibiting the multiplication of a virus with hereditary sensitivity to carbon dioxide. C. R. Hebd. Seances Acad. Sci. 236, 1085–1086.
Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59, 307–321.
Haag-Liautard, C., Dorris, M., Maside, X., Macaskill, S., Halligan, D.L., Houle, D., Charlesworth, B., Keightley, P.D., 2007. Direct estimation of per nucleotide and genomic deleterious mutation rates in Drosophila. Nature 445, 82–85.
Hense, W., Baines, J.F., Parsch, J., 2007. X chromosome inactivation during Drosophila spermatogenesis. PLoS Biol. 5, e273.
Horie, M., Honda, T., Suzuki, Y., Kobayashi, Y., Daito, T., Oshida, T., Ikuta, K., Jern, P., Gojobori, T., Coffin, J.M., Tomonaga, K., 2010. Endogenous non-retroviral RNA virus elements in mammalian genomes. Nature 463, 84–87.
Katoh, K., Misawa, K., Kuma, K., Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30, 3059–3066.
Katzourakis, A., Gifford, R.J., 2010. Endogenous viral elements in animal genomes. PLoS Genet. 6, e1001191.
Kemkemer, C., Hense, W., Parsch, J., 2011. Fine-scale analysis of X chromosome inactivation in the male germ line of *Drosophila melanogaster*. Mol. Biol. Evol. 28, 1561–1563.
Kuzmin, I.V., Novella, I.S., Dietzgen, R.G., Padhi, A., Rupprecht, C.E., 2009. The rhabdoviruses: biodiversity, phylogenetics, and evolution. Infect. Genet. Evol. 9, 541–553.
L'Heritier, P., 1958. The hereditary virus of Drosophila. Adv. Virus Res. 5, 195–245.
Li, J., Rahmeh, A., Brusic, V., Whelan, S.P., 2009. Opposing effects of inhibiting cap addition and cap methylation on polyadenylation during vesicular stomatitis virus mRNA synthesis. J. Virol. 83, 1930–1940.
Liu, H., Fu, Y., Jiang, D., Li, G., Xie, J., Cheng, J., Peng, Y., Ghabrial, S.A., Yi, X., 2010. Widespread horizontal gene transfer from double-stranded RNA viruses to eukaryotic nuclear genomes. J. Virol. 84, 11876–11887.

Longdon, B., Obbard, D.J., Jiggins, F.M., 2010. Sigma viruses from three species of Drosophila form a major new clade in the rhabdovirus phylogeny. Proc. Biol. Sci. 277, 35–44.

Longdon, B., Wilfert, L., Obbard, D.J., Jiggins, F.M., 2011a. Rhabdoviruses in two species of Drosophila: vertical transmission and a recent sweep. Genetics 188, 141–150.

Longdon, B., Wilfert, L., Osei-Poku, J., Cagney, H., Obbard, D.J., Jiggins, F.M., 2011b. Host-switching by a vertically transmitted rhabdovirus in Drosophila. Biol. Lett. 7, 747–750.

Malik, H.S., Eickbush, T.H., 1999. Modular evolution of the integrase domain in the Ty3/Gypsy class of LTR retrotransposons. J. Virol. 73, 5186–5190.

Malik, H.S., Henikoff, S., Eickbush, T.H., 2000. Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. Genome Res. 10, 1307–1318.

Miller, M.A., Pfeiffer, W., Schwartz, T., 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: Proceedings of the Gateway Computing Environments Workshop (GCE), New Orleans, LA, pp. 1–8.

Penn, O., Privman, E., Landan, G., Graur, D., Pupko, T., 2010. An alignment confidence score capturing robustness to guide tree uncertainty. Mol. Biol. Evol. 27, 1759–1767.

Stamatakis, A., Hoover, P., Rougemont, J., 2008. A rapid bootstrap algorithm for the RAxML Web servers. Syst. Biol. 57, 758–771.

Taylor, D.J., Bruenn, J., 2009. The evolution of novel fungal genes from non-retroviral RNA viruses. BMC Biol. 7, 88.

Taylor, D.J., Leach, R.W., Bruenn, J., 2010. Filoviruses are ancient and integrated into mammalian genomes. BMC Evol. Biol. 10, 193.

Taylor, D.J., Dittmar, K., Ballinger, M.J., Bruenn, J.A., 2011. Evolutionary maintenance of filovirus-like genes in bat genomes. BMC Evol. Biol. 11, 336.

Tekes, G., Rahmeh, A.A., Whelan, S.P., 2011. A freeze frame view of vesicular stomatitis virus transcription defines a minimal length of RNA for 5′ processing. PLoS Pathog. 7, e1002073.

Tordo, N., Benmansour, A., Calisher, C., Dietzgen, R.G., Fang, R.-X., Jackson, A.O., Kurath, G., Nadin-Davis, S., Tesh, R.B., Walker, P.J., 2005. Rhabdoviridae. In: Fauquet, C.M., Mayo, M.A., Maniloff, J., Desselberger, U., Ball, L.A. (Eds.), Virus Taxonomy, VIIIth Report of the ICTV, London, pp. 623–644.

Tudor, T., Davis, A.J., Feldman, M., Grammatikaki, M., O'Hare, K., 2001. The X element, a novel LINE transposable element from *Drosophila melanogaster*. Mol. Genet. Genomics 265, 489–496.

Vibranovski, M.D., Lopes, H.F., Karr, T.L., Long, M., 2009a. Stage-specific expression profiling of Drosophila spermatogenesis suggests that meiotic sex chromosome inactivation drives genomic relocation of testis-expressed genes. PLoS Genet. 5, e1000731.

Vibranovski, M.D., Zhang, Y., Long, M., 2009b. General gene movement off the X chromosome in the Drosophila genus. Genome Res. 19, 897–903.

Vicoso, B., Charlesworth, B., 2006. Evolution on the X chromosome: unusual patterns and processes. Nat. Rev. Genet. 7, 645–653.

Walker, P.J., Dietzgen, R.G., Joubert, D.A., Blasdell, K.R., 2011. Rhabdovirus accessory genes. Virus Res. 162, 110–125.

Wyers, F., Dru, P., Simonet, B., Contamine, D., 1993. Immunological cross-reactions and interactions between the Drosophila-melanogaster Ref(2)P-protein and sigma rhabdovirus proteins. J. Virol. 67, 3208–3216.

Xu, Z., Wang, H., 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 35, W265–268.

Zhang, A.B., Luo, A., Qiao, H.J., Zhang, Y.Z., Shi, W.F., Ho, S.Y.W., Xu, W.J., Zhu, C.D., 2010. Performance of criteria for selecting evolutionary models in phylogenetics: a comprehensive study based on simulated datasets. BMC Evol. Biol. 10.

Zhao, C., Dave, V., Yang, F., Scarborough, T., Ma, J., 2000. Target selectivity of bicoid is dependent on nonconsensus site recognition and protein–protein interaction. Mol. Cell. Biol. 20, 8112–8123.