Fossil rhabdoviral sequences integrated into arthropod genomes: ontogeny, evolution, and potential functionality.

Philippe Fort, Aurélie Albertini, Aurélie Van-Hua, Arnaud Berthomieu, Stéphane Roche, Frédéric Delsuc, Nicole Pasteur, Pierre Capy, Yves Gaudin, Mylène Weill

Fossil Rhabdoviral Sequences Integrated into Arthropod Genomes: Ontogeny, Evolution, and Potential Functionality

Philippe Fort,*†[1] Aurélie Albertini,†[2] Aurélie Van-Hua,[3] Arnaud Berthomieu,[4] Stéphane Roche,[2] Frédéric Delsuc,[4] Nicole Pasteur,[4] Pierre Capy,[3] Yves Gaudin,[2] and Mylène Weill*[4]

[1]Centre de Recherche de Biochimie Macromoléculaire, UMR 5237, CNRS, Universités Montpellier 2 et 1, Montpellier, France

[2]Laboratoire de Virologie Moléculaire et Structurale, Centre de Recherche de Gif, UPR 3296, CNRS, Gif sur Yvette, France

[3]Laboratoire Evolution, Centre de Recherche de Gif, Génomes et Spéciation, UPR 9034, CNRS, Gif sur Yvette, France

[4]Institut des Sciences de l'Evolution, UMR 5554, CNRS, Université Montpellier 2, C.C. 065, Place Eugène Bataillon, Montpellier, France

†These authors contributed equally to this work.

***Corresponding author:** E-mail: philippe.fort@crbm.cnrs.fr ; mylene.weill@univ-montp2.fr.

Associate editor: Manolo Gouy

Research article

Abstract

Retroelements represent a considerable fraction of many eukaryotic genomes and are considered major drives for adaptive genetic innovations. Recent discoveries showed that despite not normally using DNA intermediates like retroviruses do, Mononegaviruses (i.e., viruses with nonsegmented, negative-sense RNA genomes) can integrate gene fragments into the genomes of their hosts. This was shown for Bornaviridae and Filoviridae, the sequences of which have been found integrated into the germ line cells of many vertebrate hosts. Here, we show that Rhabdoviridae sequences, the major Mononegavirales family, have integrated only into the genomes of arthropod species. We identified 185 integrated rhabdoviral elements (IREs) coding for nucleoproteins, glycoproteins, or RNA-dependent RNA polymerases; they were mostly found in the genomes of the mosquito *Aedes aegypti* and the blacklegged tick *Ixodes scapularis*. Phylogenetic analyses showed that most IREs in *A. aegypti* derived from multiple independent integration events. Since RNA viruses are submitted to much higher substitution rates as compared with their hosts, IREs thus represent fossil traces of the diversity of extinct Rhabdoviruses. Furthermore, analyses of orthologous IREs in *A. aegypti* field mosquitoes sampled worldwide identified an integrated polymerase IRE fragment that appeared under purifying selection within several million years, which supports a functional role in the host's biology. These results show that *A. aegypti* was subjected to repeated Rhabdovirus infectious episodes during its evolution history, which led to the accumulation of many integrated sequences. They also suggest that like retroviruses, integrated rhabdoviral sequences may participate actively in the evolution of their hosts.

Key words: rhabdoviruses, evolution, inserted sequences, arthropods, *Aedes aegypti*, genomes.

Introduction

Paleovirology concerns the study of remnant viral genes found in the genomes of their hosts as well as the impact of ancient viral infections on the evolution of host genomes (Emerman and Malik 2010). Retroviruses are the best known paleoviruses, due to their high occurrence in eukaryotic genomes and the knowledge we have of their biological impact on genome dynamics (Katzourakis et al. 2005). Retroviruses are RNA viruses whose replication requires integration as proviral DNA into the genome of infected cells and can be vertically transmitted when integrated into the genome of germ line cells as endogenous retroviruses (ERVs). The future of ERVs in the host populations depends on their outcome on the host's fitness. Deleterious ERVs are expected to be rapidly lost. Neutral ERVs are submitted to genetic drift and may thus persist for much longer times; however, they will accumulate random mutations until they cannot be identified any longer. By contrast, some ERVs might carry beneficial functions for their hosts, a process known as exaptation. Classical examples are transcriptional regulation of several host genes by retroviral long terminal repeats (Buzdin et al. 2006) or resistance to superinfection conferred by the expression of envelope glycoproteins (Bishop et al. 2001). At a larger evolution scale, ERV envelope glycoproteins play a major role in fusion of trophoblasts to syncytiotrophoblasts and placenta development through their fusogenic and immunosuppressive activities (Mi et al. 2000; Mangeney et al. 2007).

An interesting consequence of virus integration in the host genome is sequence preservation. Indeed, most viruses exhibit substitution rates as high as 10^{-3} per site per round of replication for RNA viruses (reviewed in Duffy et al. 2008), which prevents analysis of their long-term evolution. However, since the host's genome evolves several orders of magnitude slower, integration therefore leads to the "fossilization" of the viral sequences.

Recently, integrated sequences derived from nonretroviral RNA viruses have been identified. As RNA viruses do not use a DNA intermediate, they were not expected to integrate into other genomes. However, sequences derived from positive strand RNA genomes were found integrated into plant and insect DNA (Crochu et al. 2004; Tanne and Sela 2005; Maori et al. 2007), and recent reports described the presence of Mononegavirales sequences integrated into vertebrate genomes (Belyi et al. 2010; Horie et al. 2010; Katzourakis and Gifford 2010; Taylor et al. 2010). Mononegavirales is an order of enveloped viruses having a nonsegmented negative-sense RNA genome (i.e., complementary to the mRNA). Mononegavirales includes four families, that is, Bornaviridae (Borna disease virus), Rhabdoviridae (e.g., Rabies virus), Filoviridae (e.g., Ebola virus), and Paramyxoviridae (e.g., measles virus), whose genomes share the same organization and sequential transcription of their five common genes: the nucleoprotein (N), a cofactor of the viral polymerase (often called P), the matrix protein (often called M), one or two glycoproteins (often called G), and the RNA-dependent RNA polymerase (L or RdRp). Integrated sequences found in the genomes of 48 vertebrate species (Belyi et al. 2010) all originated from Bornaviridae and Filoviridae, while no sequence from Rhabdoviridae could be detected despite the fact that they infect vertebrates.

Here, we searched for Rhabdoviridae sequences integrated into animal genomes. The prototypical and best-studied rhabdoviruses are the vesicular stomatitis virus (VSV), a member of the *Vesiculovirus* genus, and the rabies virus (RV), a member of the *Lyssavirus* genus. Other genera of the family include *Novirhabdovirus*, *Ephemerovirus*, *Cytorhabdovirus*, *Nucleorhabdovirus*, and the newly proposed *Dichorhabdovirus* (Kondo et al. 2006; Kuzmin et al. 2009). We detected numerous rhabdovirus-like sequences, only in the genomes of arthropod species. By using a combination of genome and field population analyses we examined their ontogeny, their evolutionary history and their potential biological impact.

Materials and Methods

Database Searches

Blastp and tblastn searches in nr, reference genomic sequences, reference mRNA and expressed sequence tag (EST) databases were performed using the NCBI Blast suite (http://blast.ncbi.nlm.nih.gov/). Specific searches were performed using VectorBase (http://www.vectorbase.org/ for *Anopheles gambiae*, *Culex quinquefasciatus*, *Aedes aegypti*, *Ixodes scapularis*, *Rhodnius prolixus*, and *Pediculus humanus*), BeetleBase (http://beetlebase.org/, *Tribolium castaneum*) and euGenes/arthropods (http://arthropods.eugenes.org/, *Acyrthosiphon pisum*, *Daphnia pulex*, *Nasonia vitripennis*). We used protein sequences of Mononegavirales from different families as queries (as indicated in supplementary table S1, Supplementary Material online).

Correspondence between ESTs and genomic loci of *A. aegypti* was deduced using VectorBase tools. DW984426 was identified by searching unassembled genomic reads (trace archives), which produced the following hits: gnl|ti|593016913, gnl|ti|263515337, gnl|ti|754271313, and gnl|ti|585847457.

Phylogenetic Inference and Molecular Clock Tests

Deduced virus-like integrated protein sequences were aligned with subsets of Rhabdoviral (N and G) or Mononegavirales (L) proteins using the MAFFT program (Katoh et al. 2002). Although many IREs contained in frame stop codons or frameshifts, we considered their full coding capacity by elimination of the mutational event. Alignments were then cleaned up using Gblocks (Talavera and Castresana 2007).

Phylogenetic trees were inferred using Bayesian inference with MrBayes v3.12 (Huelsenbeck and Ronquist 2001) and maximum likelihood (ML) with PhyML (Guindon et al. 2010). We used MrBayes with the rtREV matrix of amino acid substitution (Dimmic et al. 2002) and a gamma distribution describing among-site rate variation with eight categories (+G8). Four incrementally heated MCMCMC chains were run for 1 million generations with a sample frequency of 1,000 and a 10% burn-in value. For ML analyses, we also used the rtREV + G8 in PhyML while searching for the ML tree by performing both NNI and SPR topological moves on a bioNJ starting tree. The statistical robustness of inferred nodes was assessed by 100 bootstrap pseudoreplicates of the same ML search.

Genealogy from the 12 mitochondrial COI nucleotide sequences (848 nucleotide sites) was inferred using MrBayes under the GTR + G8 + I model. For estimation of divergence times, a likelihood ratio test for the molecular clock test was performed by comparing the ML value for the given topology with and without the molecular clock constraints under General Time Reversible model (+G8 + I) in MEGA4 (Tamura et al. 2007). The null hypothesis of equal evolutionary rate throughout the tree was not rejected at a 5% significance level ($P = 0.24$). The number of nucleotide substitutions per site was computed under a global molecular clock with a Jukes–Cantor model and a gamma distribution (shape parameter = 1). Standard error estimates were obtained by a bootstrap procedure (50 replicates). Divergence dates were then estimated by calibrating our tree with the *A. aegypti*/*A. albopictus* node estimated as 59 ± 19 My (Reidenbach et al. 2009). These molecular clock analyses were conducted with MEGA4 (Tamura et al. 2007).

d*N*/d*S* Ratio Calculations

Nonsynonymous versus synonymous substitution ratios ($\omega = d_N/d_S$) were calculated using PAML 4.4 (Yang 2007). To determine whether the ω ratios of internal and terminal branches differ significantly, we used a "one-ratio" model to estimate a single ω ratio for the entire tree and a "two-ratio" model to estimate distinct ω values for the internal branches and for the terminal branches that link duplicated IREs. The two models were then assessed by using the likelihood ratio test. d_N/d_S ratio of substitutions between Madagascar and Liverpool L42 polymerase DNA sequences of *A. aegypti* was calculated using the *KaKs* calculator program (Zhang et al. 2006).

PCR Analysis of Mosquito Samples

Aedes aegypti samples were previously described (Mousson et al. 2005). DNA was extracted from individual mosquitoes using a CetylTrimethylAmmonium Bromide (CTAB) protocol. Other samples of the *Aedes* and *Ochlerotatus* genera had been described elsewhere (Weill et al. 2004). Amplification conditions were: 3 min at 94 °C, followed by 30 cycles of 94 °C for 30 s, 52 °C for 30 s, and 72 °C for 1.5 min. Sequences were performed directly on purified products on an ABI Prism 3130 sequencer using the BigDye Terminator Kit (Applied Biosystems).

Primers used for the polymerase chain reaction (PCR) reactions were as follows: L and G fragments including 5′ nuclear *Aedes* genomic boundaries were amplified using Ldir1 (5′-GGCTGCAGCTGAGTTTGAAT-3′)/Lrev (5′-GAAAGTCCATGTGGCTTGGT-3′) and Gdir1 (5′-AGTCAGTTGTGTGGCTATGC-3′)/Grev (5′-TATCCCTCTCTCGCCACAGA-3′), respectively. L and G open reading frame (ORF) fragments were amplified using Ldir2 (5′-TGCTGAATTCGACAGAGCAG-3′)/Lrev and Gdir2 (5′-TCCATGTCGACCCGTACAGCCA-3′)/G rev, respectively. Mitochondrial cytochrome oxidase I (COI) was amplified using AeCOIF (5′-TATCGCCTAAACTTCAGCC-3′)/AeCOIR (5′-CCTAAATTTGCTCATGTTGCC-3′). Acetylcholinesterase1 (ace-1) gene PCR was performed using Ae-ace1dir (5′-CACCACTATCCGAAGACTG-3′)/Ae-ace1rev (5′-TCYAGRGTAGCAGTACC-3′) to certify DNA quality.

Sequences have been deposited in the GenBank database (HQ688271–98).

Results and Discussion

Rhabdoviridae-Like Sequences in Invertebrates Genomes

To identify integrated rhabdoviral elements (IREs), we first searched the GenBank annotated protein database using protein sequences encoded by rhabdoviruses of different genera as queries (supplementary table S1, Supplementary Material online). We identified 39 hits, including 34 N (nucleoproteins), 2 G (glycoproteins), and 3 L (RNA-dependent RNA Polymerases, RdRp) proteins (table 1). We found no hit for matrix proteins (M) or phosphoproteins (P), a feature that may result from the small sizes of these proteins and their low levels of conservation among rhabdoviruses. With the exception of a unique sequence found in zebrafish, all hits concerned invertebrates, mainly arthropods; 29 putative proteins were found in dipterans (27 in *A. aegypti*, 1 in *C. quinquefasciatus*, and 1 in *Drosophila sechellia*), 4 in arachnids (black-legged tick *I. scapularis*), 2 in nematodes (*Brugia malayi*), and 1 in zebrafish. We also identified rhabdoviral-like sequences in copepods (*Caligus rogercresseyi* and *Lepeophtheirus salmonis*) and the ant *Camponotus floridanus*. To gain in sensitivity, we next searched genome assemblies and raw sequence data directly (supplementary table S2, Supplementary Material online). We detected a much higher number of hits in *A. aegypti* (59 hits for N, 46 hits for G, and 7 for L proteins, distributed in 80 genomic contigs) and *I. scapularis* (32 hits distributed in 29 contigs). We also detected unique N- or L-related IREs in other arthropod species (several *Drosophila* species and the pea aphid *A. pisum*), as well as in the nematode *B. malayi*. We confirmed the presence of a unique N-related sequence integrated into the genome of *C. quinquefasciatus* (annotated as XP_001851570) and the absence of any N-, G-, or L-related sequences in the genome of *A. gambiae*. While this work was in progress, a general survey of endogenous viral elements in animal genomes reported IREs in insect genomes (Katzourakis and Gifford 2010). Our screening process identified twice as many IREs in *A. aegypti* and *I. scapularis* genomes and extended the analysis to all arthropod databases available so far.

To complete the identification of IREs, we searched for rhabdoviral-like sequences in EST databases (table 2): we identified 29 independent mRNAs all in arthropod databases (20 N, 4 G, and 5 L coding mRNAs), among which three in *A. aegypti* that corresponded to IREs identified previously (supercont 1.59 and 1.95 for DV253401 and DV259616 N mRNAs, respectively and unassembled raw sequences for the DW984426 G mRNA, see Materials and Methods). DV402431 was covered by 25 ESTs, which indicates that IREs can be efficiently transcribed in *A. aegypti*. By contrast, we found no ESTs derived from IREs in *I. scapularis* ESTs. This does not result from a sampling bias, since the *I. scapularis* EST database contains 193,985 sequences (vs. 303,980 for *A. aegypti* in which we found 31 ESTs for Rhabdovirus-like proteins), but rather suggests that the integrated copies are inactive in *I. scapularis*.

Searches for sequences related to other Mononegavirales families in all database types produced no significant hits in invertebrates. This contrasts with vertebrates, in particular mammals, whose genomes contain Filovirus- and Bornavirus-like but no Rhabdovirus-like integrated sequences (Belyi et al. 2010; Horie et al. 2010; Taylor et al. 2010).

Distinct IRE Integration Events in Arthropod Genomes

We next wished to examine the ontogeny of IREs in invertebrates. To this aim, we performed phylogenetic analyses of protein sequences derived from IREs and from extant Mononegavirales families (Rhabdoviridae for N and G, all families for L). We restricted our analysis to peptides larger than 200 amino acids in length. For all arthropod species, N-coding IREs clustered with *Vesiculovirus*/*Ephemerovirus* proteins. None of these IREs were related to *Lyssavirus* (fig. 1A). G-coding IRE proteins clustered with the same families, including *Lyssavirus* (fig. 1B). N and G protein sequences from viral-like copepod ESTs were as distantly related to extant Rhabdoviridae genera as to *Ixodes* or *Aedes* clusters. This supports the notion that copepod ESTs derive from genomic IREs and do not derive from current infectious viruses. We also found that in each arthropod species, N- and G-coding IREs formed well-defined clusters with closely related members at branch ends. This suggests either that a limited

Table 1. Annotated Proteins Similar to Rhabdoviral Proteins in Nonviral Databases.

Protein[a]	Accession	Description	Query Coverage[b] (%)	*E* Value[c]
N*	ACO12126.1	Nucleoprotein [*Lepeophtheirus salmonis*]	83	1.00×10^{-17}
N	XP_001660188.1	Nucleoprotein, putative [*Aedes aegypti*] > gb|EAT38605.1	87	1.00×10^{-42}
N	XP_001655472.1	Nucleoprotein, putative [*Aedes aegypti*] > gb|EAT46298.1	93	5.00×10^{-45}
N	XP_001657655.1	Nucleoprotein, putative [*Aedes aegypti*] > gb|EAT48846.1	93	5.00×10^{-45}
N	XP_001651275.1	Nucleoprotein, putative [*Aedes aegypti*] > gb|EAT48101.1	65	1.00×10^{-36}
N	XP_002411105.1	Nucleoprotein, putative [*Ixodes scapularis*] > gb|EEC13229.1	57	4.00×10^{-36}
N	XP_002435827.1	Nucleoprotein, putative [*Ixodes scapularis*] > gb|EEC08657.1	57	4.00×10^{-39}
N	XP_001660480.1	Hypothetical protein AaeL_AAEL009940 [*Aedes aegypti*]	72	1.00×10^{-29}
N	XP_001660479.1	Hypothetical protein AaeL_AAEL009870 [*Aedes aegypti*]	93	8.00×10^{-35}
N	XP_001658410.1	Nucleoprotein, putative [*Aedes aegypti*] > gb|EAT40777.1	94	3.00×10^{-40}
N	XP_001662983.1	Nucleoprotein, putative [*Aedes aegypti*] > gb|EAT45713.1	58	1.00×10^{-27}
N	XP_001657740.1	Nucleoprotein, putative [*Aedes aegypti*] > ref|XP_001657741.1	84	8.00×10^{-37}
N	XP_001652860.1	Hypothetical protein AaeL_AAEL001267 [*Aedes aegypti*]	66	3.00×10^{-26}
N	XP_001660478.1	Hypothetical protein AaeL_AAEL009873 [*Aedes aegypti*]	93	3.00×10^{-35}
N	XP_001655470.1	Nucleoprotein, putative [*Aedes aegypti*] > ref|XP_001655471.1	65	2.00×10^{-27}
N	XP_001652816.1	Nucleoprotein, putative [*Aedes aegypti*] > gb|EAT40766.1	68	3.00×10^{-29}
N	XP_001657739.1	Nucleoprotein, putative [*Aedes aegypti*] > gb|EAT47934.1	84	1.00×10^{-36}
N	XP_001650106.1	Nucleoprotein, putative [*Aedes aegypti*] > gb|EAT43638.1	75	2.00×10^{-35}
N	XP_001662846.1	Nucleoprotein, putative [*Aedes aegypti*] > gb|EAT35081.1	83	3.00×10^{-39}
N	XP_001657660.1	Hypothetical protein AaeL_AAEL000120 [*Aedes aegypti*]	65	7.00×10^{-28}
N	XP_002434822.1	Hypothetical protein IscW_ISCW005747 [*Ixodes scapularis*]	22	4.00×10^{-13}
N	XP_001851570.1	Nucleoprotein [*Culex quinquefasciatus*] > gb|EDS33713.1	84	3.00×10^{-20}
N	XP_001657738.1	Hypothetical protein AaeL_AAEL000991 [*Aedes aegypti*]	82	2.00×10^{-35}
N	XP_001655879.1	Nucleoprotein, putative [*Aedes aegypti*] > gb|EAT46028.1	68	8.00×10^{-27}
N	XP_001655880.1	Nucleoprotein, putative [*Aedes aegypti*] > gb|EAT46029.1	58	1.00×10^{-23}
N	XP_001657602.1	Hypothetical protein AaeL_AAEL006217 [*Aedes aegypti*]	58	3.00×10^{-28}
N	XP_001657601.1	Hypothetical protein AaeL_AAEL006218 [*Aedes aegypti*]	58	4.00×10^{-28}
N	XP_001655881.1	Nucleoprotein, putative [*Aedes aegypti*] > gb|EAT46030.1	58	4.00×10^{-24}
N	XP_001657665.1	Nucleoprotein, putative [*Aedes aegypti*] > gb|EAT48856.1	75	2.00×10^{-30}
N	XP_002045634.1	GM16215 [*Drosophila sechellia*] > gb|EDW54668.1	50	3.00×10^{-13}
N	XP_002402013.1	Nucleoprotein, putative [*Ixodes scapularis*] > gb|EEC02889.1	57	1.00×10^{-24}
N	XP_002412896.1	Hypothetical protein [*Ixodes scapularis*] > gb|EEC16192.1	38	3.00×10^{-12}
N	XP_001896842.1	Hypothetical protein [*Brugia malayi*] > gb|EDP34304.1	31	2.00×10^{-13}
N	XP_001650902.1	Nucleoprotein, putative [*Aedes aegypti*] > gb|EAT43064.1	55	6.00×10^{-16}
N	XP_001896395.1	Rhabdovirus nucleocapsid protein [*Brugia malayi*] > gb|EDP34757.1	61	1.00×10^{-17}
G	XP_001658709.1	Hypothetical protein AaeL_AAEL007844 [*Aedes aegypti*]	41	1.00×10^{-07}
G*	ACO10239.1	Glycoprotein G precursor [*Caligus rogercresseyi*]	43	4.00×10^{-07}
L*	EFN64915.1	Large structural protein [*Camponotus floridanus*]	10	1.00×10^{-19}
L	XP_001653948.1	Hypothetical protein AaeL_AAEL001772 [*Aedes aegypti*]	28	2.00×10^{-19}
L	XP_002666898.1	Hypothetical protein [*Danio rerio*]	12	2.00×10^{-03}

[a] Annotated proteins derive from genomic sequences, except those labeled with an asterisk, identified from RNA EST.

[b] Query coverage corresponds to the fraction of the query protein covered by the annotated protein.

[c] Only hits with *E* values below e^{-01} are shown.

number of founder IREs spread by multiple duplication events or that multiple integrations of closely related viruses took place in a limited period of time. The same question applies to the clustering of the *C. quinquefasciatus* IRE (CPIJ010057, arrow in fig. 1A) with the *A. aegypti* cluster N3; either sequences from closely related viruses have entered the genome of each species independently or a single integration event predated the *Aedes*–*Culex* split (116 to 217 Ma, Krzywinski et al. 2006; Reidenbach et al. 2009).

L-coding IREs encoded polymerase fragments (118–923 amino acids of over 2,000 in rhabdoviruses), which in *A. aegypti* did not all overlap each other. We thus performed two independent analyses for the N- and C-moieties (fig. 1C and *D*), which revealed a more complex situation than that of N- and G-coding IREs; The nine *I. scapularis* copies formed a monophyletic cluster related to *Vesiculovirus*/*Ephemerovirus*/*Lyssavirus*, whereas the five L-coding IREs in *A. aegypti* were clearly polyphyletic, which supports the scenario of independent integration events. Furthermore, the five IREs appear distantly related, as illustrated by Aae_L59 and Aael_L1077, grouped with *Vesiculovirus*/*Ephemerovirus*/*Lyssavirus* and Aael_L42, as distantly related from Rhabdoviridae as from Filoviridae. Aael_L42 and the unique IRE found in the zebrafish genome form a cluster distinct from the Midway/Nyamanini virus group, contrarily to what was reported recently (Belyi et al. 2010).

In conclusion, N-, G-, and L-coding IREs in arthropods are polyphyletic and do not cluster with sequences from extant *Rhabdovirus* genera. Furthermore, L-coding IREs are related to different Mononegavirales families. These data suggest that IREs originated from different viral populations and are consistent with the occurrence of several integration episodes.

Mechanisms of Integration of Rhabdovirus-Like Sequences

In Mononegavirales, genes are transcribed from the RNA genome as monocistronic mRNAs (Villarreal et al. 1976).

Table 2. Rhabdoviridae-Related Proteins Identified in Nonviral RNA EST Databases.

Protein	Species	*E* Value[a]	EST Accession[b]
G	*Aedes aegypti*	2.00×10^{-04}	*DW987726*; EG005928
G	*Caligus clemensi*	1.50×10^{-02}	GO408331
G	*Caligus rogercresseyi*	8.00×10^{-05}	*ACO10239*; FK886840; FK881621; FK888518
G	*Lepeophtheirus salmonis*	5.00×10^{-12}	EY508221

N	*Aedes aegypti*	5.00×10^{-20}	*DV402431*; DV245753; DV249095; DV402431;
			DV254100; DV427558; DW219075;
			DV245751; DV246409; EB092583;
			DV246423; DV246407; DV249093; DV246421;
			DV266277; DV266276;
			DV398395; DV296797; DW992445; DV430094;
			DV411933; DV233375; DV349257; DV246409; DV253401

N	*Aedes aegypti*	8.00×10^{-19}	*DV253916*; DV238796; DV244810; DV238795
N	*Boophilus microplus*	9.00×10^{-23}	FG302076
N	*Caligus clemensi*	3.00×10^{-10}	*GO404700*; GO404701
N	*Caligus rogercresseyi*	2.00×10^{-13}	*FK881980*; FK881981
N	*Caligus rogercresseyi*	4.00×10^{-31}	*FK880615*; FK880616
N	*Caligus rogercresseyi*	3.00×10^{-25}	*FK898446*; FK898447
N	*Cimex lectularius*	1.00×10^{-10}	GR909184
N	*Diabrotica virgifera*	3.00×10^{-07}	EW766035
N	*Lepeophtheirus salmonis*	2.00×10^{-21}	*FK912858*; FK912859
N	*Lepeophtheirus salmonis*	8.00×10^{-20}	HO697345
N	*Lepeophtheirus salmonis*	3.00×10^{-14}	*EX482495*; EX482496
N	*Lepeophtheirus salmonis*	1.00×10^{-11}	*GW629304*; FK926026; FK926027
N	*Lepeophtheirus salmonis*	7.00×10^{-11}	*EX485667*; EX485668
N	*Lepeophtheirus salmonis*	1.00×10^{-10}	*EX480043*; EX480044
N	*Lepeophtheirus salmonis*	8.00×10^{-09}	FK916520
N	*Lutzomyia longipalpis*	2.00×10^{-14}	*AM092396*; AM092394
N	*Spodoptera littoralis*	2.00×10^{-15}	GW825907
N	*Spodoptera littoralis*	4.00×10^{-07}	FQ019503

N	*Tetranychus urticae*	9.00×10^{-18}	*GW058809*; GW007518; GW016127; GW016126; GW021367;
			GW017077; GW018748;
			GW021366; GW058787; GW026927

L	*Caligus rogercresseyi*	2.00×10^{-81}	*FK899171*; FK899172
L	*Caligus rogercresseyi*	1.00×10^{-67}	*FK887655*; FK870029; FK887100
L	*Lepeophtheirus salmonis*	2.00×10^{-27}	*EY507182*; EY507187
L	*Lepeophtheirus salmonis*	1.00×10^{-11}	HO689963
L	*Heliothis virescens*	0.054	GT055020

[a] *E* values correspond to the lowest values from all performed TBLASTN searches.

[b] Accession numbers used as references in trees shown in figure 1 are in italics.

IREs from *A. aegypti* and other species encoded only three protein types (N, G, or L), which were never found associated as would have been expected from integration of a genomic template. This supports the notion that IREs were primarily generated through reverse transcription of viral mRNAs. This probably involved the machinery of transposable elements (TEs), since the number of IREs appears to correlate positively with the genomic TE content of their hosts. Indeed, nearly 50% of the *A. aegypti* genome consists of TEs (Nene et al. 2007) versus 29% for *C. quinquefasciatus* (Arensburger et al. 2010) and only 16% for *A. gambiae* (Kaufman et al. 2002). A potential role of TEs in *A. aegypti* also agrees with the high occurrence of polyA tracts downstream of IREs (indicated by asterisks in supplementary table S2, Supplementary Material online). This situation could also pertain to *I. scapularis* IREs, since the tick genome contains many TEs, in particular a high content of short interspersed elements (SINEs), which reflects an intense retroposition activity (Sunter et al. 2008). In mammals, propagation of SINEs was recently put forward as the major actor of de novo integration of Bornaviridae sequences in living cells (Horie et al. 2010).

Retrotransposition of Mononegavirales mRNAs was nevertheless proposed to be a rare event that requires specific conditions, as suggested by the low number of integrated copies in vertebrates (Taylor et al. 2010). Such low integration efficacy is consistent with the absence of endogenous rhabdovirus or paramyxovirus sequences in vertebrate genomes or the absence of Sigma virus IRE in the *Drosophila* genome, whereas this virus stably infects *Drosophila* germ line cells (Longdon et al. 2011). On the other hand, the presence of multiple IREs found in a few arthropod species also suggests that under favorable conditions, retrotransposition can be an efficient process.

Besides retrotransposition, we identified in each cluster sequences showing hallmarks of recent duplications: high level of DNA sequence similarity, presence of homologous non-viral 5′ and 3′ flanking sequences (not shown) or arrangement in tandem (see supplementary table S2, Supplementary Material online). As duplicated IREs only differ by mutations accumulated since the duplication event, we computed substitution rates and estimated when the eldest duplications occurred (table 3). We assumed comparable mutation rates of nuclear genes in mosquitoes and

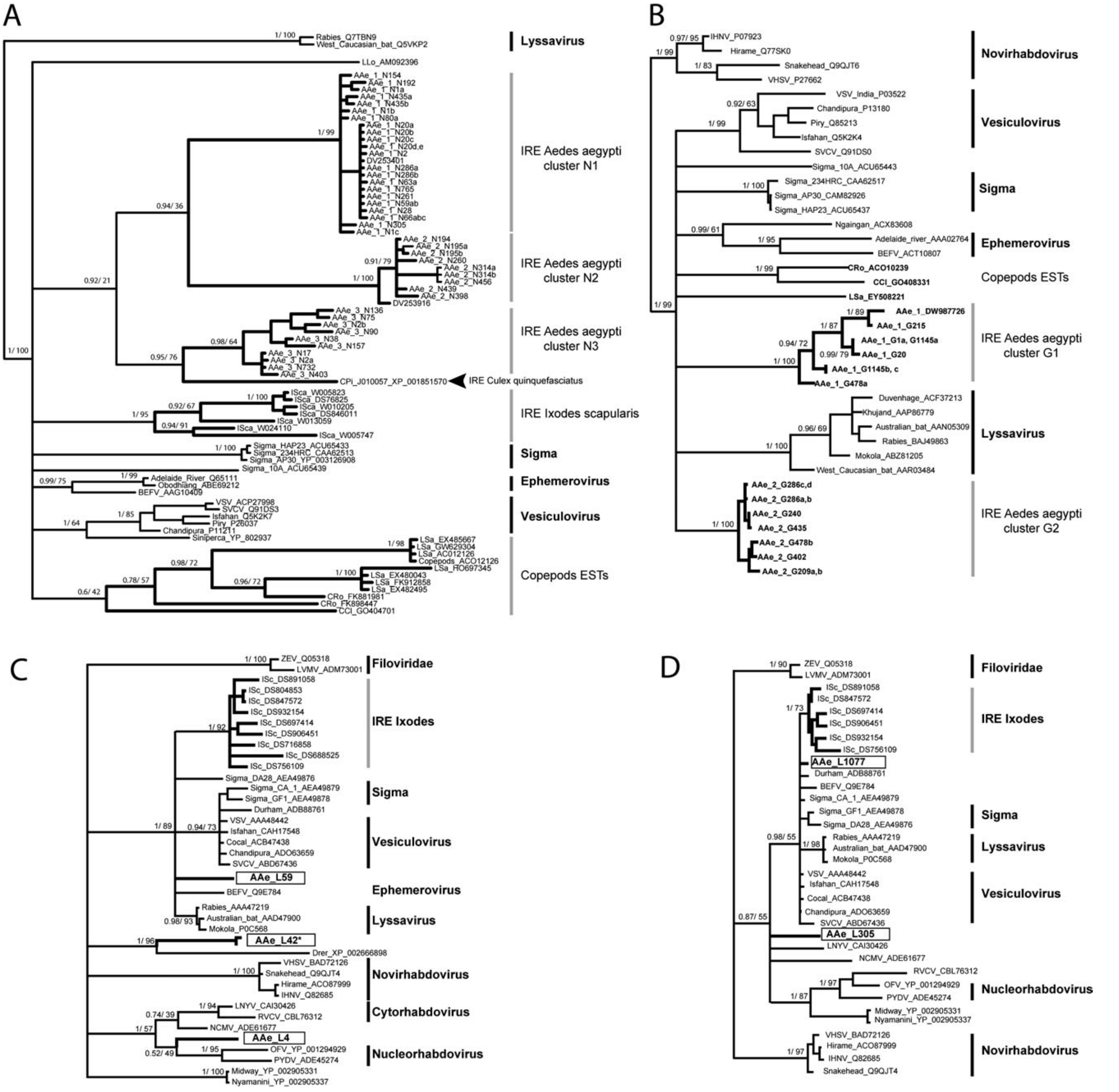

Fig. 1. Clustering of invertebrate IREs and rhabdoviral proteins. Phylogenetic trees were inferred from alignments of protein sequences deduced from IREs, ESTs, and Mononegavirales proteins. Nucleoproteins (*A*), glycoproteins (*B*), polymerases: N-terminal (first 1,000 amino acids) (*C*), and C-terminal (*D*) *Aedes aegypti* IREs have thicker branches. Trees are 50% majority consensus obtained from Bayesian inference. Statistical supports (Bayesian posterior probabilities and ML bootstrap values) are indicated above branches (omitted for clarity in terminal branches). The L42 ortholog of Malagasy *A. aegypti* samples is indicated by an asterisk in (*C*).

Drosophila, that is, 3.5 × 10^{-9} per site per generation (Keightley et al. 2009) and an average generation time of 27 days for *A. aegypti* (Yakob et al. 2008). The highest observed substitution rate (cluster G, 0.073) suggests that the eldest duplications occurred 1.5 Ma.

IREs in the *A. aegypti* Genome Derived from Numerous Integration Events

Topologies of the trees shown in figure 1 indicate that the five L-coding IREs integrated independently into the *A. aegypti* genome while N- and G-coding IREs clusters were probably generated by retrotranspositions of viral mRNAs combined with genomic duplications. Retrotransposed IREs differ by mutations accumulated since their integration and by the divergence between the founder rhabdoviral sequences. Assuming that ancestral RNA viruses had mutation rates similar to extant ones, that is, 10^{-3} per site per year (Duffy et al. 2008), variability between founder rhabdoviruses is expected to prevail. Retrotransposed IREs are thus expected to display traces of past purifying selection (d_N/d_S

Table 3. Dating of Duplication Events.

Clusters	Duplicated IREs	Substitution Rate Per Site[a]	Duplication Age[b]
N1	N1.20a, N1.20b N1.20c N1.20d N1.20e	0.002 ± 0.001	45.8 ± 15.9
	N1.59a, N1.59b	0.001	21.1
	N1.66a, N1.66b N1.66c	0.002 ± 0.001	42.3 ± 21.1
	N1.154, N1.305	0.023	486.1
	N1.286a, N1.286b	0.003	63.4
	N1.435a, N1.435b	0.013	274.8
N2	N1.195a, N1.195b	0.002	42.3
	N1.314a, N1.314bN1.456	0.000	<20
N3	N1.2a, N1.732	0.002	42.3
	G1.209a, G1.209b	0.073	1542.9
G1	G1.286a, G1.286b	0.005	105.7
	G1.286c, G1.286d	0.017	369.7
G2	G1.1145a, G1.1a G1.20	0.014 ± 0.002	295.9 ± 36.6
	G1.1145b, G1.1145c	0.011	232.5

[a] Substitution rates were computed with MEGA4 using a Maximum Composite Likelihood model.

[b] Duplication age was estimated using a 3.5 × 10^{-9} nuclear substitution rate per year and a 27-day generation time for *A. aegypti*. Ages are expressed as thousand years.

ratios < 1, i.e., ratios of nonsynonymous to synonymous changes). By contrast, most IREs generated by gene duplication should behave as neutral sequences (d_N/d_S ratios ≈ 1).

We found very low global d_N/d_S ratio values for N- and G-coding IREs (0.00846 and 0.02696, respectively, one-ratio, table 4). Since IREs are unlikely to all be under purifying selection, this result strongly suggests that most IREs were generated by independent integration events. To strengthen this observation, we used the two-ratio model, in which internal and terminal branches can have different d_N/d_S ratio values. We found much higher d_N/d_S ratio values on terminal branches that connect duplicated IREs identified in table 3 (0.763 and 0.344 for N- and G-coding IREs, respectively), in agreement with a neutral behavior.

Taken together, these data support a scenario wherein 44 IREs (68%) sharing low d_N/d_S ratio were generated by independent integrations of related viral sequences, whereas the remaining 20 with higher d_N/d_S ratio were generated by duplications. This indicates that *A. aegypti* was subjected to repeated Rhabdovirus infectious episodes during its evolutionary history, which led to the accumulation of viral sequences integrated independently.

Rhabdoviral-Like G and L Sequences in Worldwide *A. aegypti* Field Populations

Aedes aegypti IREs were only identified in a single reference genome (the Liverpool strain). We next examined the presence and variability of IREs in nine *A. aegypti* field populations sampled worldwide. We focused on the two annotated proteins—the G286a and L42 proteins—encoded by the longest integrated ORFs. Despite the presence of repeated sequences, we successfully designed specific PCR primer sets that allowed to detect the two IREs in most *A. aegypti* samples (supplementary table S3, Supplementary Material online). G286a sequences were highly conserved, displaying only 14 variable sites over 1 kb (supplementary fig. S1*A*, Supplementary Material online). L42 PCR fragments were shorter in half of the samples compared with the Liverpool reference strain, due to the absence of a 619-bp DNA sequence 5′ upstream of the viral ORF (supplementary fig. S1*B*, Supplementary Material online). Aside from this variable insert, L42 sequences only differed from each other by 1–3 nucleotide substitutions. The G286a and L42 IREs thus display low levels of variability and appear to be nearly fixed in worldwide *A. aegypti* populations.

We next examined more distant species of the *Aedes* and *Ochlerotatus* genera. Attempts failed to detect the two IREs in species other than *A. aegypti* (supplementary table S4, Supplementary Material online). Whether these IREs are indeed missing or were not detected because of mutations in the amplimer targets remains to be clarified. We analyzed *A. aegypti* isolated from Indian Oceanic islands, in which population variation was previously examined (Failloux et al. 2002). We found low levels of variability in two individuals from Europa Island

Table 4. Analysis of d_N/d_S Ratios of N- and G-Coding IREs in *Aedes aegypti* Samples.

Cluster[a]	One-Ratio Whole Tree	Two-Ratio Internal Branches	Two-Ratio Terminal Branches (duplications)	*P* Value[b]
N-coding	0.0085	0.0064	0.7636	<0.001
N1	0.0126	0.0081	0.4834	<0.001
N2	0.0031	0.0016	1.2876	<0.01
N3	0.0105	0.0036	n.a.[c]	n.a.
G-coding	0.0270	0.0250	0.3448	<0.01
G1	0.0245	0.0111	0.1798	<0.05
G2	0.0151	0.0138	0.3165	<0.01

[a] Clusters refer to figure 1.

[b] *P* value was estimated from the χ^2 distribution.

[c] Not applicable (single duplication with identical IREs).

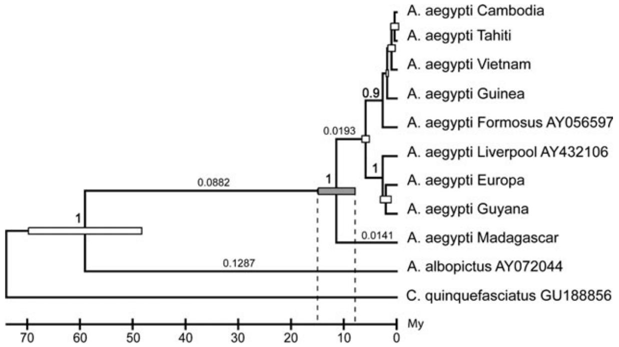

Fig. 2. Divergence time of *Aedes aegypti* sampled in Madagascar. Phylogenetic inference of Malagasy and other *A. aegypti* samples was constructed from a multiple alignment of mitochondrial COI DNA sequences using MrBayes. For clarity, only posterior probabilities above 0.9 are shown (in bold below branches). Substitution rates inferred under a global molecular clock are shown on above branches. Divergence times and associated standard errors were obtained by calibrating the tree on the *A. aegypti*/*A. albopictus* node using the molecular clock program (MEGA4 software) and are expressed as My.

(supplementary fig. S1 and table S3, Supplementary Material online). We also identified L42 sequences in two Malagasy individuals (supplementary table S3, Supplementary Material online). Although divergent from the other *A. aegypti* L42 sequences (supplementary fig. S2A, Supplementary Material online), the Malagasy sequences were clearly L42 orthologues, since 1) they branched together, whereas the remaining four L IREs delineated distinct stems (see fig. 1C), 2) they encoded partial polymerases truncated at the same amino acid positions, and 3) they showed identical boundaries of viral and nonviral sequences.

Variation between Malagasy and mainland L42 sequences suggested that the Malagasy samples may belong to a distinct *A. aegypti* clade, whose divergence postdated integration of the L42 IRE. To address the time at which the Malagasy clade emerged, we analyzed a 1,107-bp long fragment of the mitochondrial cytochrome oxidase I (COI) in two individuals. The two Malagasy sequences were identical to each other but differed from the reference Liverpool strain by 38 variable sites (3.43%). Phylogenetic analysis confirmed that the Malagasy mtDNA formed a specific *A. aegypti* branch (fig. 2), in agreement with a previous analysis (Mousson et al. 2005). Using a standard mtDNA clock estimated in insects (2.3–3.54% My-1, Brower 1994; Papadopoulou et al. 2010), we found that the mtDNA of the Malagasy clade emerged 1–1.5 Ma. However, the standard mitochondrial calibration may be overestimated for mosquito lineages (Krzywinski et al. 2006). In particular, it would place the *A. aegypti*/*A. albopictus* node at 3–5 My only, whereas it was estimated at 59 ± 19 My from analysis of six nuclear genes (Reidenbach et al. 2009). Using this value as a calibration constraint on the *A. aegypti*/*A. albopictus* node, the mtDNA of the Malagasy clade might be as old as 11.4 ± 4.1 Ma (fig. 2), that is, in the range of the estimated ages of the N- or G-coding IRE duplications (see table 3). In conclusion, although for diverse reasons, analysis of mtDNA alone may produce misleading conclusions on the phylogenetic history of species (Galtier et al. 2009), our data suggest that the L42 IRE integrated several million years ago, supporting the notion that Rhabdoviridae is an ancient family of the Mononegavirales order, like Bornaviruses and Filoviruses, already present 10–40 Ma (Belyi et al. 2010; Taylor et al. 2010).

Probable Exaptation of a Rhabdoviral Polymerase Fragment in *A. aegypti*

A major aspect of IRE integration concerns their potential role in the biology of their hosts. IREs detrimental to their hosts are expected to be rapidly lost mainly through allele segregation. Neutral IREs can spread in populations by genetic drift but are expected to accumulate random mutations and degenerate until they can no longer be identified as IREs. By contrast, IREs that confer an increased fitness to their hosts are expected to spread throughout the host populations and show higher sequence conservation. Although IREs identified in this study encode only protein fragments and many of them display frameshifts or in-frame stop codons, they can still have a positive impact on the metabolism of their hosts, such as a protective effect against viral infections through the synthesis of dominant negative fragments or antisense RNAs that may perturb the viral replication cycle. Due to the high noise introduced by independent integrations, a d_N/d_S ratio-based approach to detect IREs that may confer an advantageous phenotype can only be applied to orthologous IREs. We focused on the L42 sequence, present in Malagasy and other *A. aegypti* populations (supplementary fig. S2, Supplementary Material online). L42 sequences appeared to be under purifying selection, as demonstrated by the low d_N/d_S ratio (below 0.044 whatever the substitution model used, P value $< 2.4 \times 10^{-62}$, supplementary table S5, Supplementary Material online). This strongly suggests that the polymerase sequence has been exapted by the *Aedes* genome.

Exaptation might also concern nucleoprotein IREs, some of which have kept full-length coding capacities. However, to be addressed, this requires either an estimate of the time at which IRE sequences became integrated, which is not yet available, or the identification of orthologous IREs in other distantly related isolates, like Malagasy *A. aegypti*. From a functional perspective, IREs can contribute to immunity against subsequent infections by the same or related viruses. Expression of viral-like proteins or protein fragments may indeed exert dominant negative effects on processes critical for the virus cycle, such as nucleocapsid assembly, which requires multimerization of N proteins (Albertini et al. 2006; Green et al. 2006) or assembly of transcription and replication complexes. In addition to antiviral immunity, these sequences may have been exapted for other functions, base on their ability to bind and protect RNA.

Conclusions

Our data demonstrate that rhabdoviral sequences were present at least several millions years ago and repeatedly entered the genomes of a few arthropod species. Integration probably requires very particular conditions, which may explain the absence of rhabdovirus and paramyxovirus sequences in vertebrate genomes or the absence of sigma virus sequences in the *Drosophila* genome, whereas this virus stably infects the germ cell line. Our data also show that over half of the IREs found in *A. aegypti* originated from independent integration events, thereby constituting a fossil snapshot of the diversity present in ancestral viral populations. Last, we identified a potential case of exaptation by analyzing *A. aegypti* field populations.

Additional sampling of related *A. aegypti* populations is now necessary to examine the contribution of IREs to the selective fitness of their hosts, such as increased immunity described in vertebrates and potentially to their vector efficacy.

Supplementary Material

Supplementary figures S1–S2 and tables S1–S5 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Anna-Bella Failloux for providing the *Aedes aegypti* samples. We thank Julian Venables for critical reading the manuscript. All sequence data were obtained on the Environmental Genomic Platform of the SFR Montpellier-Environnement-Biodiversité. This work was supported by CNRS institutional grants and by an ANR contract (ANR-08-BLAN-0256), including a postdoctoral fellowship to A. Albertini. Contribution 2011-124 of the Institut des Sciences de l'Evolution de Montpellier (UMR CNRS 5554).

References

Albertini AA, Wernimont AK, Muziol T, Ravelli RB, Clapier CR, Schoehn G, Weissenhorn W, Ruigrok RW. 2006. Crystal structure of the rabies virus nucleoprotein-RNA complex. *Science* 313:360–363.

Arensburger P, Megy K, Waterhouse RM, et al. (76 co-authors). 2010. Sequencing of Culex quinquefasciatus establishes a platform for mosquito comparative genomics. *Science* 330:86–88.

Belyi VA, Levine AJ, Skalka AM. 2010. Unexpected inheritance: multiple integrations of ancient bornavirus and ebolavirus/marburgvirus sequences in vertebrate genomes. *PLoS Pathog.* 6:e1001030.

Bishop KN, Bock M, Towers G, Stoye JP. 2001. Identification of the regions of Fv1 necessary for murine leukemia virus restriction. *J Virol.* 75:5182–5188.

Brower AV. 1994. Rapid morphological radiation and convergence among races of the butterfly Heliconius erato inferred from patterns of mitochondrial DNA evolution. *Proc Natl Acad Sci U S A.* 91:6491–6495.

Buzdin A, Kovalskaya-Alexandrova E, Gogvadze E, Sverdlov E. 2006. At least 50% of human-specific HERV-K (HML-2) long terminal repeats serve in vivo as active promoters for host nonrepetitive DNA transcription. *J Virol.* 80:10752–10762.

Crochu S, Cook S, Attoui H, Charrel RN, De Chesse R, Belhouchet M, Lemasson JJ, 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.

Dimmic MW, Rest JS, Mindell DP, Goldstein RA. 2002. rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. *J Mol Evol.* 55:65–73.

Duffy S, Shackelton LA, Holmes EC. 2008. Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev Genet.* 9:267–276.

Emerman M, Malik HS. 2010. Paleovirology—modern consequences of ancient viruses. *PLoS Biol.* 8:e1000301.

Failloux AB, Vazeille M, Rodhain F. 2002. Geographic genetic variation in populations of the dengue virus vector Aedes aegypti. *J Mol Evol.* 55:653–663.

Galtier N, Nabholz B, Glemin S, Hurst GD. 2009. Mitochondrial DNA as a marker of molecular diversity: a reappraisal. *Mol Ecol.* 18:4541–4550.

Green TJ, Zhang X, Wertz GW, Luo M. 2006. Structure of the vesicular stomatitis virus nucleoprotein-RNA complex. *Science* 313:357–360.

Guindon S, Dufayard JF, 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.

Horie M, Honda T, Suzuki Y, et al. (11 co-authors). 2010. Endogenous non-retroviral RNA virus elements in mammalian genomes. *Nature* 463:84–87.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754–755.

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. 2010. Endogenous viral elements in animal genomes. *PLoS Genetics.* 6:e1001191.

Katzourakis A, Rambaut A, Pybus OG. 2005. The evolutionary dynamics of endogenous retroviruses. *Trends Microbiol.* 13:463–468.

Kaufman TC, Severson DW, Robinson GE. 2002. The Anopheles genome and comparative insect genomics. *Science* 298:97–98.

Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S, Blaxter ML. 2009. Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines. *Genome Res.* 19:1195–1201.

Kondo H, Maeda T, Shirako Y, Tamada T. 2006. Orchid fleck virus is a rhabdovirus with an unusual bipartite genome. *J Gen Virol.* 87:2413–2421.

Krzywinski J, Grushko OG, Besansky NJ. 2006. Analysis of the complete mitochondrial DNA from Anopheles funestus: an improved dipteran mitochondrial genome annotation and a temporal dimension of mosquito evolution. *Mol Phylogenet Evol.* 39:417–423.

Kuzmin IV, Novella IS, Dietzgen RG, Padhi A, Rupprecht CE. 2009. The rhabdoviruses: biodiversity, phylogenetics, and evolution. *Infect Genet Evol.* 9:541–553.

Longdon B, Wilfert L, Osei-Poku J, Cagney H, Obbard DJ, Jiggins FM. 2011. Host-switching by a vertically transmitted rhabdovirus in Drosophila. *Biol Lett.* 7:747–750.

Mangeney M, Renard M, Schlecht-Louf G, Bouallaga I, Heidmann O, Letzelter C, Richaud A, Ducos B, Heidmann T. 2007. Placental syncytins: genetic disjunction between the fusogenic and immunosuppressive activity of retroviral envelope proteins. *Proc Natl Acad Sci U S A.* 104:20534–20539.

Maori E, Tanne E, Sela I. 2007. Reciprocal sequence exchange between non-retro viruses and hosts leading to the appearance of new host phenotypes. *Virology* 362:342–349.

Mi S, Lee X, Li X, et al. (12 co-authors). 2000. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* 403:785–789.

Mousson L, Dauga C, Garrigues T, Schaffner F, Vazeille M, Failloux AB. 2005. Phylogeography of Aedes (Stegomyia) aegypti (L.) and Aedes (Stegomyia) albopictus (Skuse) (Diptera: Culicidae) based on mitochondrial DNA variations. *Genet Res.* 86:1–11.

Nene V, Wortman JR, Lawson D, et al. (95 co-authors). 2007. Genome sequence of Aedes aegypti, a major arbovirus vector. *Science* 316:1718–1723.

Papadopoulou A, Anastasiou I, Vogler AP. 2010. Revisiting the insect mitochondrial molecular clock: the mid-Aegean trench calibration. *Mol Biol Evol.* 27:1659–1672.

Reidenbach KR, Cook S, Bertone MA, Harbach RE, Wiegmann BM, Besansky NJ. 2009. Phylogenetic analysis and temporal diversification of mosquitoes (Diptera: Culicidae) based on nuclear genes and morphology. *BMC Evol Biol.* 9:298.

Sunter JD, Patel SP, Skilton RA, Githaka N, Knowles DP, Scoles GA, Nene V, de Villiers E, Bishop RP. 2008. A novel SINE family occurs frequently in both genomic DNA and transcribed sequences in ixodid ticks of the arthropod sub-phylum Chelicerata. *Gene* 415:13–22.

Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol.* 56:564–577.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol Biol Evol.* 24:1596–1599.

Tanne E, Sela I. 2005. Occurrence of a DNA sequence of a non-retro RNA virus in a host plant genome and its expression: evidence for recombination between viral and host RNAs. *Virology* 332:614–622.

Taylor DJ, Leach RW, Bruenn J. 2010. Filoviruses are ancient and integrated into mammalian genomes. *BMC Evol Biol.* 10:193.

Villarreal LP, Breindl M, Holland JJ. 1976. Determination of molar ratios of vesicular stomatitis—virus induced RNA species in Bhk21 cells. *Biochemistry* 15:1663–1667.

Weill M, Berthomieu A, Berticat C, Lutfalla G, Negre V, Pasteur N, Philips A, Leonetti JP, Fort P, Raymond M. 2004. Insecticide resistance: a silent base prediction. *Curr Biol.* 14:R552–R553.

Yakob L, Alphey L, Bonsall MB. 2008. Aedes aegypti control: the concomitant role of competition, space and transgenic technologies. *J Appl Ecol.* 45:1258–1265.

Yang ZH. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.

Zhang Z, Li J, Zhao XQ, Wang J, Wong GK, Yu J. 2006. KaKs calculator: calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinform.* 4:259–263.