Endogenous RNA viruses of plants in insect genomes

Jie Cui, Edward C. Holmes

A B S T R A C T

Endogenous viral elements (EVEs) derived from RNA viruses with no DNA stage are rare, especially those where the parental viruses possess single-strand positive-sense (ssRNA+) genomes. Here we provide evidence that EVEs that share a sequence similarity to ssRNA+ viruses of plants are integrated into the genomes of a number of insects, including mosquito, fruit flies, bees, ant, silkworm, pea aphid, Monarch butterfly, and wasps. A preliminary phylogenetic analysis places these EVEs as divergent relatives of the Virgaviridae and three currently unclassified plant viral species.
and wasps (*Nasonia vitripennis*, *N. longicornis*, and *N. giraulti*) contain EVEs that exhibit the closest relationship to plant viruses (protein sequence similarities of 50%–22%; e-values 0.81–2e−39; Table S1). The reverse BLAST analysis revealed that these EVEs were most closely related to the members of the *Virgaviridae* (ssRNA + viruses) and the two unclassified plant viruses (*blueberry necrotic ring blotch virus* and *citrus leprosis virus* C; sequence similarities 44%–25%; e-values 3e−04–3e−167, Table S2). To further evaluate these potential integration events, we retrieved the positive hit contigs and obtained key information (including flanking host gene) on the presence and location of EVEs in the insect genomes (Table S1). Despite the strong sequence-based evidence for many of these EVEs, the authenticity of the short and/or fragmentary copies will require additional confirmation by deep sequencing of host genomes. No EVEs related to plant viruses were observed in the 74 chordeate genomes.

This analysis also revealed distant hits to members of the genus *Alphavirus* (family *Togaviridae*), a group of insect (predominantly mosquito) transmitted viruses of animals (*Forrester et al.*, 2011). Of these, the closest match was with *Semliki forest virus* (GenBank accession CA770503), although relatively weak compared to most of the matches involving the plant viruses described above (query replicase protein of EVEaa, e-value = 2e−10; Identities = 82/334 = 25%). Hence, this distant match to the alphaviruses may simply reflect the fact that the RNA-dependent RNA polymerase is the most conserved protein in all RNA viruses, as suggested by our phylogenetic analysis (see below).

To determine the genomic structure of one of these novel EVEs, we retrieved contig1.29253 (GenBank accession AA02029253) of the *A. aegypti* (mosquito) genome and located sequences clearly homologous to each of the three major virus genes – replicase, movement protein (MP), and CP (Fig. 1) – thereby confirming their viral origin. All three genes contain multiple indels (insertions/deletions) and premature stop codons (an alignment between a representative EVE and the most closely related plant viruses is presented in Fig. S1). Not only does the divergent nature of the EVE genome strongly argue against contamination by exogenous viruses (and the relevant host and viruses were sequenced in different laboratories), but it suggests that the invasions of these plant viruses into host germ line cells were ancient events. The fragmentary nature of these integrated sequences also precluded any analysis of selection pressures.

To establish a provisional evolutionary history of these novel endogenous viral elements, we conducted a phylogenetic analysis using the replicase protein sequences of the three EVEs that exhibited the longest amino acid sequences from *A. aegypti*, *D. rhopaloa*, and *B. bombus*. The replicase protein sequences of plant viruses identified in the BLAST search (n = 22) were downloaded from GenBank (Table S3) to form a comparative set, as were those of the following representative alphaviruses (n = 6): *Semliki forest virus* (GenBank accession CA770503), *Chikungunya virus* (AD095920), *Eastern equine encephalitis virus* (ABB45867), *Venezuelan encephalitis virus* (AAD14564), *Sindbis virus* (AAC83378), and *Ockelbo virus*, (AAA96972). Amino acid sequence alignment was performed in ClustalX 2.0 (*Larkin et al.*, 2007) with manual adjustment undertaken using Se-Al (http://tree.bio.ed.ac.uk/software/seal). This resulted in an alignment of 472 amino acids. However, because of the divergent nature of these sequence data, we removed ambiguously aligned regions using the Gblocks program (*Talavera and Castresana*, 2007). This resulted in a final alignment of 185 amino acids from 31 taxa which we used as the basis of our phylogenetic analysis. Phylogenetic relationships were inferred using the maximum likelihood method in PhyML 3.0 (*Guindon et al.*, 2010), assuming the WAG + I model of amino acid substitution. Phylogenetic robustness was determined using 1000 bootstrap replicates. All *Virgaviridae* species clustered closely together in the phylogenetic tree, while the three unclassified species formed a second monophyletic group, albeit without strong bootstrap support (Fig. 2). Notably, this phylogenetic analysis placed all the insect EVEs as more closely related to the three unclassified plant viruses, although with weak bootstrap support (63%) and which likely reflects the very short sequence analyzed and its divergent nature. However, there was strong (100%) bootstrap support for the clustering of these EVEs with all the exogenous plant viruses used in this analysis and to the exclusion of the far more divergent insect-transmitted alphaviruses. Finally, it is also noteworthy that two of the EVEs (*EVEdr*, *D. rhopaloa*; and *EVEbt*, *B. terrestris*) formed a relatively well-supported monophyletic group, although with long internal branch lengths.

In sum, we have identified EVEs most closely related to ssRNA + viruses of plants in 14 insect genomes. Although we cannot entirely exclude that these insect EVEs were in fact derived from currently undocumented viruses of insect origin, particularly as so few insect viruses have been characterized to date, (i) our phylogenetic analysis clearly places these EVEs within the (known) genetic diversity of plant RNA viruses, and (ii) these EVEs seemingly contain a protein homologous to the movement protein of plants viruses suggests that they are in fact of plant origin. Clearly, the frequent interactions between plants and insects, for example through pollination and feeding, and that many plant viruses are insect-borne, may allow the occasional invasion of insect genomes by plant viruses. The *Virgaviridae* are a good case in point. This newly classified family of plant viruses (including the genera *Furovirus*, *Hordeiviridae*, *Pecuvirus*, *Pomovirus*, *Tomamovirus*, and *Tobravirus*) infects a very wide range of divergent plant hosts, including herbaceous, mono- and dicotyledonous species (*Adams et al.*, 2009). Viral transmission is occasionally mediated by insects, including aphids, leafhoppers, thrips, and whiteflies, although usually with no replication in the vector (*Adams et al.*, 2009).

Even less clear is the time-scale of these integration events. There are 21 *Drosophila* species in our screening database, although only three (*D. rhopaloa*, *D. ananassae*, and *D. ficuspilus*) were observed to harbor plant virus EVEs. However, as there is no chromosome location information for these two species, we cannot determine whether these EVEs are present as orthologous sequences. The failure to discover plant virus EVEs in the other 19 *Drosophila* genomes suggests that EVEs seen in these two genomes were not vertically transmitted, although this is clearly a question that needs to be explored in greater detail. Bees, ants, and wasps, which seemingly possess related EVEs, belong to the order *Hymenoptera* that diverged from the *Diptera* (mosquitoes and fruit flies) and *Lepidoptera* (silkworm) approximately 355 million years ago (*Wiegmann et al.*, 2009), an evolutionary time-scale that has also been proposed for some large DNA viruses of insects (*Thézé et al.*, 2011). It is therefore possible that plant viruses also invaded insect genomes on such a deep time-scale. However, resolution of this issue clearly requires further study and particularly the identification of EVEs inserted into orthologous positions in hosts with known divergence times. Interestingly, in our phylogenetic analysis, *EVEdr* (*Diptera*) grouped with EVEbt (*Hymenoptera*) to the exclusion of EVEaa (*Diptera*), tentatively suggesting multiple invasions of plant viruses into insect genomes.
Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.virol.2012.02.014.

Fig. 2. Phylogenetic relationships of insect EVEs, some ssRNA+ viruses of plants, and selected members of the genus Alphavirus (Togaviridae). Bootstrap values (>60%) are shown for key nodes and branch lengths are drawn to a scale of amino acid substitutions per site. The tree is midpoint rooted for purposes of clarity only. The three EVEs are marked in red and represent those obtained from the genomes of A. aegypti (EVEaa), D. rhopaloa (EVEdr), and B. terrestris (EVEbt).

References


