Genome sequence of a crustacean iridovirus, IIV31, isolated from the pill bug, Armadillidium vulgare

Benoit Piégu, Sébastien Guizard, Tan Yeping, Corinne Cruaud, Sassan Asgari, Dennis K. Bideshi, Brian A. Federici and Yves Bigot

The family Iridoviridae consists of large dsDNA viruses that infect species of both poikilothermic vertebrates (fishes, amphibians and reptiles) and invertebrates (arachnids, cephalopods, crustaceans, insects, molluscs, nematodes and polychaetes; Williams, 2008). These viruses are members of the order Megavirales (Colson et al., 2013), so-called nucleocyttoplasmic large DNA viruses (NCLDV; Iyer et al., 2001). The dsDNA genomes of iridoviruses are circularly permuted with terminal redundancy. As a consequence, the map of their genomes is represented as a circular molecule. Only one linear molecule is encapsidated in each virion, with the ends of individual encapsidated genomes being located at different positions on the map of different virions (Bigot et al., 2000). The genome of vertebrate iridoviruses is highly methylated, whereas little or no methylation occurs in the genomes of the invertebrate iridoviruses (IIV). Replication of the iridoviral genome includes distinct nuclear and cytoplasmic phases, but virions only assemble in the cytoplasm (Jancovich et al., 2000). The genome of vertebrate iridoviruses is highly methylated, whereas little or no methylation occurs in the genomes of the invertebrate iridoviruses (IIV). Replication of the iridoviral genome includes distinct nuclear and cytoplasmic phases, but virions only assemble in the cytoplasm (Jancovich et al., 2011). The genomes are encapsidated within an icosahedral capsid ranging between 120 and 180 nm in diameter. Capsids are composed predominantly of a 50 kDa major capsid protein. The IIV studied by cryo-electron microscopy have surface fibrils 2 nm in diameter (Yan et al., 2000).

The taxonomic structure of the family Iridoviridae is currently organized into five genera: Chlorovirus, Iridovirus, Lymphocystivirus, Megalocytivirus and Ranavirus. Members of the two first genera have a host range restricted to invertebrate species, whereas the three others infect only poikilothermic vertebrates. The model species for the genus Chlorovirus is IIV3 (Delhon et al., 2006; Jancovich et al., 2011), the only species reported in this genus. The type species for the genus Iridovirus is IIV6, and only two species, IIV1 and IIV6, have been recognized by the International Committee for Virus Taxonomy (ICTV) as representatives of this genus. Ten other related viruses that may be Iridovirus species await biological and genomic data before it can be determined whether they are valid species or variants of existing species. Interestingly, the phylogenetic analyses of proteins encoded by IIV3, IIV6 (Jakob et al., 2001) and IIV9 (Wong et al., 2011) have revealed that IIV9 is more closely related to IIV3 than to IIV6. This has been confirmed with IIV22, IIV22A, IIV25 and IIV30, four close relatives of IIV9 (Piégu et al., 2013a, b, c, d), indicating that some species of the genus Iridovirus are more closely related to members of the genus Chlorovirus than to other Iridovirus species from insects. Consequently, this suggests that the genus
Iridovirus contains several diverse species or species complexes. Division of the Iridovirus genus into three species complexes, the Polyiridovirus (type species IIV9), the Oligoiridovirus (type species IIV6) and the Crustaceoiridovirus (type species IIV31), has been proposed (Williams, 1994; Williams & Cory, 1994), among which members of the Polyiridovirus complex would share a common iridovirus ancestor with those of the genus Chloriridovirus. This proposal has not been accepted formally by the ICTV, but in agreement with it are data showing that members of the Oligoiridovirus complex are the closest IIV relatives of the family Ascoviridae (Bigot et al., 2011), and share a common IIV ancestor (Stasiak et al., 2003; Bigot et al., 2009). Moreover, ascoviruses and IIVs are more closely related to each other than to vertebrate iridoviruses. Although first determined through phylogenetic analyses, this classification is strongly supported by data showing that these invertebrate viruses share 26 core genes, of which only 19 are found in vertebrate iridoviruses (Eaton et al., 2007; Bigot et al., 2009). A precise definition of genera among the IIVs therefore remains unresolved at present, and further investigations are required to elucidate whether or not species complexes should be elevated to genus rank.

To date, six genomes of IIVs have been sequenced: IIV3, the mosquito iridescent virus (Delhon et al., 2006), IIV6, the Chilo iridescent virus (Jakob et al., 2001), IIV9, the Wiseana iridovirus (Wong et al., 2011), IIV22 (Piégu et al., 2013a), IIV22A (Piégu et al., 2013b), IIV25 (Piégu et al., 2013c) and IIV30 (Piégu et al., 2013d). Here, we present a summary classification and a set of features for IIV31, the eighth IIV sequenced, together with the description of the sequencing and annotation of its genome. To date, the classification status of IIV31 reveals that it is a tentative species belonging to the genus Iridovirus, in which it is related to the Crustaceoiridovirus complex (Williams, 1994; Williams & Cory, 1994; Jancovich et al., 2011).

IIV31 was originally isolated from a sample of adult pill bugs of the species Armadillidium vulgare (Crustacea, Isopoda, Oniscidea) in southern California on the campus of the University of California, Riverside, USA (Federici, 1980; Williams, 1994). IIV31 has been found to have a wide host range in populations of crustacean species (Wijnhoven & Berg, 1999). Large quantities of virions can be directly purified from adults recovered from natural habitats such as leaf litter. However, it can also be produced per os and by injection of healthy isopods, or larvae of a coleopteran (Cole & Morris, 1980). Here, virions of iridovirus type IIV31 were harvested in April 2010 from isopods collected from a bed of ivy on the campus of the University of California, Riverside, and frozen at −80 °C. IIV31-infected isopod adults are recognizable by a characteristic blue discoloration of the normally grey cuticle. IIV31 virions and their genomic DNA (gDNA) were purified as described elsewhere (Federici, 1980; Bigot et al., 2009). The purity of our IIV31 sample was determined by digesting virion gDNA with restriction enzymes using published data as a reference (Federici, 1980; Williams, 1994).

In 2009, the scientific committee of GENOSCOPE selected the IIV31 genome for sequencing. The complete genome sequence and annotation are now available in the EMBL database (accession no. HF920637). A summary of the project results is shown in Table 1. The genome of IIV31 was sequenced using the 454 FLX pyrosequencing platform (Roche/454). Library construction and sequencing were performed as previously described elsewhere (Henn et al., 2010). De novo genome assembly was performed using the Newbler v2.3 assembly software package as previously described elsewhere (Henn et al., 2010). Assembly metrics are described in Table 1. The assembled contig representing the entire IIV30 genome sequence was confirmed by comparing five predicted restriction fragment profiles from the genome, for BamHI, EcoRI, HindIII, PstI and SalI, with the matching fragment profiles produced by actual restriction digestions of the IIV31 genome (Federici, 1980; Williams, 1994).

Genes were identified using the Broad Institute Automated Phage Annotation Protocol as described elsewhere (Ashburner et al., 2000; Henn et al., 2010). Briefly, evidence-based and ab initio gene prediction algorithms were used to identify putative genes, followed by construction of a consensus gene model using a rules-based evidence approach. Gene models were manually checked for errors such as in-frame stops, very short peptides, splits and merges. Additional gene prediction analysis and functional annotation were performed as previously described (Bigot et al., 2009).

General features of the IIV31 genome sequence (Table 2) include a nucleotide composition of 35.09 mol% G+C (Fig. 1). Pair alignment using BLASTN of the IIV31 genome with those of the IIV3, IIV6, IIV9, IIV22, IIV22A, IIV25 and IIV30 genomes did not reveal any identity at the level of nucleotide sequences. Similarly, we did not find any conserved cluster of collinear genes between IIV31 and the other IIVs.

A total of 203 genes encoding proteins were predicted. No genes coding for tRNAs were found. Of the 203 coding DNA sequences (CDSs), 125 were in forward orientation and 78 in reverse orientation. Four gene pairs were found to overlap: 103R/104L, 114R/155L, 129L/130R and 162L/163R. Eighty-four CDSs (41.4 %) have been annotated with functional product predictions. The annotation of the 203 genes, described in Table S1 (available in the online Supplementary Material), revealed that 184 of the 203 CDSs have a related gene in databases, with e values below 10−2. One hundred and seventy-two genes have an orthologue in the IIV6 genome, three genes (004R, 040R, 067L) have an orthologue in the IIV9 genome that did not occur in the IIV6 genome, four genes (015L, 024R, 084R, 119R) have a viral orthologue that did not occur in one of the currently sequenced IIV genomes, and five genes (013L, 022L, 082L, 128L, 136L) have no orthologue in any currently sequenced viral genomes, but have an orthologue in a prokaryotic or a eukaryotic genome. Finally, 19 genes have no orthologue in databases and putatively correspond...

With regard to repeats, three families of gene paralogues occur in the IIV31 genome. The first contains 10 members that are related to CIV genes 006L, 019R, 029R, 146R, 148R, 211L, 212L, 238R, 313L, 388R, 420R and 468L. The second contains three members related to CIV261R, 396L and 443R. The third contains eight members that belong to the family bro-like genes, a widespread family of repeated genes in NCLDVs (Bideshi et al., 2003).

Three gene fossils, i.e. genes containing stop codons or frameshifts, were found in IIV31. The status of these genes was confirmed by PCR and sequencing, and therefore we decided to annotate them as fossil genes. The first is located between nucleotides 27 211 and 27 755. It is a derivative of a mimivirus gene, MIMI_R865, that belongs in the Acanthamoeba polyphaga mimivirus genome to a family of gene paralogues to which also belong genes MIMI_L17, MIMI_L298 and MIMI_L754. The second fossil is located between nucleotides 190 307 and 191 099. It was determined as a member of the family bro-like genes.

The presence of certain mobile genetic elements that occur in some NCLDVs belonging to the families Phycodnaviridae, Mimiviridae and Ascoviridae was searched for in the IIV31 genome (Desnues et al., 2012; Bao & Jurka, 2013). No transpovirons, group I introns and Fanzor1 or Fanzor2 DNA transposons were found. In contrast to IIV6, IIV9, IIV22, IIV22A, IIV25 and IIV30, no inteins were found inserted into the ORF 097R of IIV31, which encodes the alpha subunit of the ribonucleotide-diphosphate reductase. However, one intein was found to be specifically inserted in-frame in 001R, which encodes the delta DNA polymerase, as reported previously (Bigot et al., 2013).

A miniature transposon (MITE) was found between nucleotides 38 829 and 39 878. This 1049 bp MITE was named IIV31-MITE. We think that this is a derivative of a class II transposon (Wicker et al., 2007). Indeed, both of its extremities correspond to inverted terminal repeats (ITRs) about 371 bp in length that are juxtaposed at their outer ends by a CTAG tetranucleotide that corresponds to a

---

**Table 1. Genome sequencing project information**

<table>
<thead>
<tr>
<th>MIGS ID</th>
<th>Property</th>
<th>Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIGS-31</td>
<td>Finishing quality</td>
<td>Finished (&gt;99 %)</td>
</tr>
<tr>
<td></td>
<td>No. of contigs</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Assembly size</td>
<td>220 222 bp</td>
</tr>
<tr>
<td></td>
<td>Assembly coverage</td>
<td>30 x</td>
</tr>
<tr>
<td></td>
<td>Total number of reads</td>
<td>33 821</td>
</tr>
<tr>
<td>MIGS-29</td>
<td>Sequencing plate form</td>
<td>454</td>
</tr>
<tr>
<td></td>
<td>Gene calling method</td>
<td>Annotation protocol – Bigot et al. (2009)</td>
</tr>
<tr>
<td></td>
<td>EMBL accession no.</td>
<td>HF920637</td>
</tr>
</tbody>
</table>

---

**Table 2. Genome statistics of IIV31**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Value</th>
<th>% of total*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size (bp)</td>
<td>220 222</td>
<td>100</td>
</tr>
<tr>
<td>G+C content (bp)</td>
<td>77 276</td>
<td>35.09</td>
</tr>
<tr>
<td>Coding region (bp)</td>
<td>189 465</td>
<td>86.0</td>
</tr>
<tr>
<td>Pseudogenes</td>
<td>3</td>
<td>100</td>
</tr>
<tr>
<td>Total genes (putatively functional)</td>
<td>203</td>
<td>100</td>
</tr>
<tr>
<td>Protein-encoding genes with function prediction</td>
<td>84</td>
<td>41.3</td>
</tr>
<tr>
<td>Protein-encoding genes with orthologues in databases</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Family of gene paralogues</td>
<td>3</td>
<td>–</td>
</tr>
<tr>
<td>Genes in families of paralogous genes</td>
<td>21</td>
<td>10.4</td>
</tr>
<tr>
<td>Non-coding regions over 200 bp in length</td>
<td>10 788 (20 segments)</td>
<td>4.9</td>
</tr>
</tbody>
</table>

*The total is based on either the size of the genome in bp or the total number of protein-encoding genes in the annotated genome, where applicable.
target site duplication (TSD), which occurred at the time of the MITE insertion (Fig. 2). Its inner regions contain two repeat motifs that are 39 and 10 bp in length, each repeated four times. In the IIV31 genome, this MITE overlapped two small paralogous CDSs, 034R and 035L (Table S1), that had no orthologue in other sequenced iridoviruses.

The nucleotide sequence comparison of IIV31-MITE with IIV9-MITE and IIV22-MITE revealed that they were not related. Interestingly, the TSD and the ITR length of the IIV31-MITE match those of a prokaryotic family of insertion sequence, IS5. An IS5-like transposon with similar TSD and ITR features, IS5_Av, has already been described in the genome of a bdelloid rotifer, Adineta vaga (Gladyshev & Arkhipova, 2009). Here, we propose that IIV31-MITE and IS5_Av could represent the only two members of a currently unknown eukaryotic family of transposable elements that would be related to the IS5 prokaryotic family. However, no sequence data for the transposase that mediated the transposition of IIV31-MITE are available to confirm this proposition.

IIV31 is, to our knowledge, the first iridovirus genome infecting a crustacean species to be sequenced and reported. This genome revealed 19 new putative proteins with sizes varying from 105 to 435 aa residues. Many (184) of the CDSs identified displayed high conservation with their counterparts in other IIVs, insect and bacterial genomes. Further sequencing of related strains will reveal more about the genetic and functional diversity of these interesting viruses.

The discovery of IIV31-MITE, together with that of IIV09-MITE and IIV22-MITE, and of Fanzor1 and 2 DNA transposons (Bao & Jurka, 2013) in two ascoviruses, SfAV1a and HvAV3e, suggests that these closely related invertebrate viruses (Stasiak et al., 2003; Bigot et al., 2009) might be, a priori, vectors of horizontal transfers of transposable elements between host species. However, their presence in these genomes must be considered as unexpected for two reasons. First, the cellular niche of these viruses is mostly cytoplasmic, whereas that of the DNA transposons is only nuclear. Second, the genome of these viruses has a molecular configuration in which the DNA is not negatively supercoiled or is negatively supercoiled to only a small extent (Bigot et al., 2000). Such genome configurations are incompatible with some needs of DNA transposons. Indeed, DNA transposons need a negatively supercoiled DNA environment for an efficient mobility at the excision and the insertion sites (Sinzelle

---

**Fig. 1.** Circular map of the 220 222 bp IIV31 genome. The outer scale is numbered clockwise in bp. Circles 1 and 2 (from outside to inside) denote CDSs (forward strands in red and reverse strands in blue). Green boxes in circle 3 represent ORF-free regions with a size of over 200 bp. The orange boxes in circle 4 represent the three fossil genes. Circle 5 represents the local variations of G+C content along the genome sequence (green, regions with a GC content above the average; purple, regions with a GC content below the average).

**Fig. 2.** Nucleotide sequence of the IIV31-MITE that was found in the region spanning nucleotides 38 829 and 39 878. ITRs are highlighted in black with white text, and the duplicated CTGA tetranucleotide at the insertion site is in bold type. The four 39 bp motifs that are tandemly repeated in the inner region are in italic type and alternately highlighted in light or dark grey. The four TATAAAAATT motifs are underlined.
sons and MITEs are repeated elements that are interspersed based on two observations. First, DNA transposons bearing them in their genome. This might find selection pressure, which would be advantageous for the viruses bearing them in their genome. This might find support based on two observations. First, DNA transposons and MITE transcripts can assemble in intrastand dyad structures (Petit et al., 2007), i.e. short hairpin RNA (shRNA) molecules that are good substrates for the RNA interference machineries. Second, IIVs and ascoviruses seem to be transcribed during the viral cycle (D’Costa et al., 2004), our hypothesis is that shRNA transcripts from DNA transposon and MITE provide another way used by these viruses to modify the expression of some host genes by RNA interference. Further sequencing of the genome of their hosts will be required to verify the consistency of this hypothesis.

Acknowledgements
This work was supported by the C.N.R.S. and GENOSCOPE, and funded by the Ministère de l’Education Nationale, de la Recherche et de la Technologie and the Groupement de Recherche CNRS 3546.

References


