Transcription of densovirus endogenous sequences in the *Myzus persicae* genome

Gabriel Clavijo,† Manuela van Munster,‡ Baptiste Monsion,† Nicole Bochet† and Véronique Brault†

1UMR 1131 SVQV INRA-UDS, 28 rue de Herrlisheim, 68021 Colmar, France
2UMR 385 BGPI, INRA-CIRAD-SupAgro, CIRAD TA-A54/K, Campus International de Baillarguet, 34398 Montpellier, France

Integration of non-retroviral sequences in the genome of different organisms has been observed and, in some cases, a relationship of these integrations with immunity has been established. The genome of the green peach aphid, *Myzus persicae* (clone G006), was screened for densovirus-like sequence (DLS) integrations. A total of 21 DLSs localized on 10 scaffolds were retrieved that mostly shared sequence identity with two aphid-infecting viruses, *Myzus persicae* densovirus (MpDNV) and Dysaphis plantaginea densovirus (DplDNV). In some cases, uninterrupted potential ORFs corresponding to non-structural viral proteins or capsid proteins were found within DLSs identified in the aphid genome. In particular, one scaffold harboured a complete virus-like genome, while another scaffold contained two virus-like genomes in reverse orientation. Remarkably, transcription of some of these ORFs was observed in *M. persicae*, suggesting a biological effect of these viral integrations. In contrast to most of the other densoviruses identified so far that induce acute host infection, it has been reported previously that MpDNV has only a minor effect on *M. persicae* fitness, while DplDNV can even have a beneficial effect on its aphid host. This suggests that DLS integration in the *M. persicae* genome may be responsible for the latency of MpDNV infection in the aphid host.

**INTRODUCTION**

Insects are the most diverse and largest animal group, causing economically significant losses in a wide range of crops, either as direct pests or as disease vectors (Chilana et al., 2012). Studying insect biology and virus transmission by insects is therefore critical to identify control targets to minimize their impact. Advances in sequencing technology have dramatically increased the amount of information available on insect genomes, creating new opportunities for development of research programmes focused on insect biology. Deep-sequence analysis of insect genomes has revealed that they may contain integrated viral sequences that are not of retroviral origin (Katzourakis & Gifford, 2010; Liu et al., 2011b; Thezé et al., 2011). These ancestral endogenous sequences share similarity with viruses in different families: ssDNA viruses (family Paroviridae), dsDNA viruses (families Baculoviridae, Polydnaviridae and Nudiviridae), reverse-transcribed DNA viruses (family Hepadnaviridae), ssRNA positive-sense (family Flaviviridae) or negative-sense (families Rhabdoviridae and Orthomyxoviridae) viruses, and dsRNA viruses (family Reoviridae) (Bézier et al., 2009; Cheng et al., 2014; Katzourakis & Gifford, 2010; Liu et al., 2011c; Thezé et al., 2011). Integration of such a great diversity of viral sequences suggests that many different replication strategies can induce integration of these sequences into host genomes. Typically, viral sequences are found scattered throughout insect genomes, and in most cases, no complete virus genome has been found. However, in some cases, integrated viral ORFs are uninterrupted and can therefore potentially be expressed in host cells (Bézier et al., 2009; Katzourakis & Gifford, 2010; Liu et al., 2011b). In plants, insects and vertebrates, integration of viral sequences is predicted to confer selective advantages to the host by generating antiviral immunity against related viruses (Belyi et al., 2010; Bertsch et al., 2009; Flegel, 2009; Goic et al., 2013; Maori et al., 2007).

Viral sequences from viruses in the subfamily Densovirinae (family Paroviridae) have been shown to be integrated into the genome of *Drosophila*, *Rhodnius prolixus* and the pea aphid (*Acyrthosiphon pisum*) (Liu et al., 2011b). Densoviruses (DNVs) are small, non-enveloped ssDNA
viruses encapsidated in icosahedral particles of about 25 nm diameter. Their genomes are 4–6 kb and encode only a few proteins referred to as either non-structural (NS) or capsid (CP, also referred to as VP) proteins (Bergoin & Tijssen, 2010). NS proteins are implicated in virus replication and in the activation of the promoter driving the synthesis of CP proteins (Legendre & Rommelaere, 1992; Rhode & Richard, 1987), while CP proteins are involved in virion formation (Crozier et al., 2000). Genome organization varies among the different genera in the subfamily, with the two sets of genes being in tandem on the same DNA strand in the genera *Iteravirus*, *Brevidensovirus*, *Hepandensovirus* and *Penstyl densovirus* and on complementary strands in the genus *Ambidensovirus*. DNVs are found in a wide range of insect families and can cause acute diseases that are often fatal for their insect host (Barreau et al., 1996; Mutuel et al., 2010; Suto, 1979; Szélei et al., 2011). As DNVs are not known to infect mammals (Dupont, 2003; El-Far et al., 2004), biopesticides derived from engineered or natural DNVs have been developed to reduce lepidopteran, mosquito and cockroach populations (Buchatski et al., 1987; Carlson et al., 2006; Hu et al., 2000; Jiang et al., 2007; Monserrat et al., 1984; Tal & Attathom, 1993). However, some DNVs may only induce sublethal infection and may even confer a beneficial effect on their host (i) by increasing the production of the winged morph and therefore promoting insect dispersal, or (ii) by protecting against pathogenic baculovirus infection (Ryabov et al., 2009; van Munster et al., 2003a; Xu et al., 2014).

We utilized the recent release of the genome sequence of the aphid *Myzus persicae* to look for integrated viral sequences originating from viruses of the subfamily *Densovirinae*. We identified and confirmed the presence of genomic sequences sharing partial homology with *Myzus persicae* densovirus (MpDNV) and *Dysaphis plantaginea* densovirus (DpIDNV) (Ryabov et al., 2009; van Munster et al., 2003b). Interestingly, we also showed that these integrated viral sequences are transcribed in *M. persicae*. Previous studies have shown that MpDNV has a moderate effect on the development and fecundity of *M. persicae* with no significant effect on the intrinsic growth rate of the population (van Munster et al., 2003a). It is possible that integration and transcription of densovirus-like sequences (DLSs) sharing similarities with MpDNV could be responsible for the low pathogenicity of MpDNV in *M. persicae*.

**RESULTS**

**Detection and organization of DLSs in the *M. persicae* genome**

We screened the assembled genome of *M. persicae* (clone G006) for integrated sequences homologous to DNV, hereafter referred to as DLSs. BLAST searches were performed using as queries the reference strain sequences from each of the five genera in the subfamily *Densovirinae* (*Brevidensovirus*, *Hepandensovirus*, *Iteradensovirus*, *Penstyl densovirus* and *Ambidensovirus*) together with the genome sequences of the two known aphid DNVs, MpDNV and DpIDNV. Twenty-one genomic sequences sharing significant sequence identity with MpDNV and DpIDNV (E-values ranging from 2.10^{-11} to 1.10^{-2}) were identified on 10 scaffolds (Table 1). An additional genomic sequence from scaffold_883 exhibited sequence identity with the lepidopteran densovirus Junonia coenia densovirus (JcDNV), the type member of the genus *Ambidensovirus* (Table 1). DLS integrations covered genomic sequence lengths ranging from 87 to 1887 bp. Embedded in these sequences, deduced amino acid sequences of different lengths (from 29 to 664 aa) shared 24–74 % identity with NS or CP proteins of MpDNV, DpIDNV or JcDNV (Table 1).

An almost-complete ambidensovirus-like genome was found on scaffold_1839 with the NS protein-coding sequences on one strand of the genome and the CP protein-coding sequences on the complementary strand. This organization of the coding sequences was consistent with that of genomes found in infectious virions (Fig. 1). A more peculiar situation was encountered on scaffold_794 where two almost-complete ambidensovirus-like genomes were integrated, both in reverse orientation (Fig. 1). This coding sequence structure may imply two independent but neighbouring insertions in the genome. The other scaffolds contained ambidensovirus-like sequences sharing homology with either NS or CP protein sequences. In most cases, these viral sequences were only partial. As several integrations were detected at the beginning or at the end of a scaffold (scaffold_564, _1826, _1646, _1268 and _1885), we cannot completely exclude the existence of additional virus-like sequences that may be missing in the available data.

The presence of DLS integrations in *M. persicae* genome was confirmed in another available *M. persicae* genome sequence (clone_O). All the integration events detected in the G006 clone were observed in clone_O (Table S1, available in the online Supplementary Material). This analysis was extended to the genome of another aphid species, *A. pisum* (International Aphid Genomics Consortium, 2010). Using an updated full-genome sequence of *A. pisum* (assembly v2), we retrieved 53 DLS integrations (Table S2). Previously, Liu et al. (2011b) described the presence of 49 DLS integrations in the *A. pisum* genome. The discrepancy between our analyses and those of Liu et al. (2011b) might result from the new assembly of an *A. pisum* genome. In contrast to the *M. persicae* genome, the DLS integrations in *A. pisum* shared sequence identity not only with MpDNV, DpIDNV and JcDNV but also with two other DNVs belonging to the genus *Ambidenso virus*, *Galleria mellonella* densovirus (GmDNV) and Helicoverpa armigera densovirus (HaDNV) (Table S2). The observed DLS integrations covered genomic sequence lengths ranging from 165 to 462 bp. Portions of these sequences (55–1547 aa) shared amino acid sequence identity ranging from 23 to 46% when compared with the
Table 1. Inventory of DLS integrations within the *M. persicae* genome (clone G006) using five genera of DNVs as queries

<table>
<thead>
<tr>
<th>Scaffold no.</th>
<th>Scaffold size (bp)</th>
<th>Nt position of integration</th>
<th>DNV genus</th>
<th>Top viral BLAST hit (reference strain)</th>
<th>Related gene</th>
<th>GenBank accession no.</th>
<th>Aa ID (%)</th>
<th>Alignment length (aa)</th>
<th>E-value</th>
<th>Name of DLS in Phylogeny</th>
</tr>
</thead>
<tbody>
<tr>
<td>scaffold_28</td>
<td>1079708</td>
<td>674429–675373</td>
<td><em>Ambidensovirus</em> Dysaphis plantaginea densovirus</td>
<td>NS</td>
<td>gil204324151</td>
<td>gb</td>
<td>ACI01073_1</td>
<td>43.0</td>
<td>316</td>
<td>2E–76</td>
</tr>
<tr>
<td>scaffold_28</td>
<td>1079708</td>
<td>673450–673863</td>
<td><em>Ambidensovirus</em> Dysaphis plantaginea densovirus</td>
<td>NS</td>
<td>gil204324151</td>
<td>gb</td>
<td>ACI01073_1</td>
<td>24.0</td>
<td>138</td>
<td>1E–08</td>
</tr>
<tr>
<td>scaffold_413</td>
<td>266 173</td>
<td>193157–192810</td>
<td><em>Ambidensovirus</em> Myzus persicae densovirus</td>
<td>CP</td>
<td>gil33235701</td>
<td>ref</td>
<td>NP_874378_1</td>
<td>42.2</td>
<td>116</td>
<td>2E–25</td>
</tr>
<tr>
<td>scaffold_564</td>
<td>193 318</td>
<td>1124–1231</td>
<td><em>Ambidensovirus</em> Dysaphis plantaginea densovirus</td>
<td>CP</td>
<td>gil204324153</td>
<td>gb</td>
<td>ACI01075_1</td>
<td>58.3</td>
<td>36</td>
<td>1E–04</td>
</tr>
<tr>
<td>scaffold_794</td>
<td>108 655</td>
<td>11516–12106</td>
<td><em>Ambidensovirus</em> Myzus persicae densovirus</td>
<td>CP</td>
<td>gil33235702</td>
<td>ref</td>
<td>NP_874377_1</td>
<td>33.0</td>
<td>200</td>
<td>5E–24</td>
</tr>
<tr>
<td>scaffold_794</td>
<td>108 655</td>
<td>13426–12422</td>
<td><em>Ambidensovirus</em> Myzus persicae densovirus</td>
<td>NS</td>
<td>gil33235700</td>
<td>ref</td>
<td>NP_874376_1</td>
<td>51.2</td>
<td>365</td>
<td>3E–99</td>
</tr>
<tr>
<td>scaffold_794</td>
<td>108 655</td>
<td>13904–13491</td>
<td><em>Ambidensovirus</em> Myzus persicae densovirus</td>
<td>NS</td>
<td>gil33235699</td>
<td>ref</td>
<td>NP_874375_1</td>
<td>36.4</td>
<td>143</td>
<td>3E–14</td>
</tr>
<tr>
<td>scaffold_794</td>
<td>108 655</td>
<td>15297–15734</td>
<td><em>Ambidensovirus</em> Myzus persicae densovirus</td>
<td>NS</td>
<td>gil33235699</td>
<td>ref</td>
<td>NP_874375_1</td>
<td>35.8</td>
<td>151</td>
<td>1E–14</td>
</tr>
<tr>
<td>scaffold_794</td>
<td>108 655</td>
<td>15799–16803</td>
<td><em>Ambidensovirus</em> Myzus persicae densovirus</td>
<td>NS</td>
<td>gil33235700</td>
<td>ref</td>
<td>NP_874376_1</td>
<td>51.2</td>
<td>365</td>
<td>3E–99</td>
</tr>
<tr>
<td>scaffold_794</td>
<td>108 655</td>
<td>18741–17119</td>
<td><em>Ambidensovirus</em> Myzus persicae densovirus</td>
<td>CP</td>
<td>gil33235702</td>
<td>ref</td>
<td>NP_874377_1</td>
<td>41.1</td>
<td>560</td>
<td>1E–108</td>
</tr>
<tr>
<td>scaffold_794</td>
<td>108 655</td>
<td>18750–18601</td>
<td><em>Ambidensovirus</em> Dysaphis plantaginea densovirus</td>
<td>CP</td>
<td>gil204324153</td>
<td>gb</td>
<td>ACI01075_1</td>
<td>74.00</td>
<td>50</td>
<td>5E–18</td>
</tr>
<tr>
<td>scaffold_794</td>
<td>108 655</td>
<td>19148–18762</td>
<td><em>Ambidensovirus</em> Dysaphis plantaginea densovirus</td>
<td>CP</td>
<td>gil204324154</td>
<td>gb</td>
<td>ACI01076_1</td>
<td>41.9</td>
<td>136</td>
<td>2E–20</td>
</tr>
<tr>
<td>scaffold_883</td>
<td>69 518</td>
<td>62012–62188</td>
<td><em>Ambidensovirus</em> Junonia coenia densovirus</td>
<td>CP</td>
<td>gil257676</td>
<td>gb</td>
<td>AAB23698_1</td>
<td>39.0</td>
<td>59</td>
<td>3E–04</td>
</tr>
<tr>
<td>scaffold_1268</td>
<td>27 708</td>
<td>714–370</td>
<td><em>Ambidensovirus</em> Dysaphis plantaginea densovirus</td>
<td>CP</td>
<td>gil204324154</td>
<td>gb</td>
<td>ACI01076_1</td>
<td>60.9</td>
<td>115</td>
<td>1E–34</td>
</tr>
<tr>
<td>scaffold_1335</td>
<td>26 965</td>
<td>15334–15789</td>
<td><em>Ambidensovirus</em> Dysaphis plantaginea densovirus</td>
<td>CP</td>
<td>gil204324154</td>
<td>gb</td>
<td>ACI01076_1</td>
<td>40.3</td>
<td>154</td>
<td>1E–22</td>
</tr>
<tr>
<td>scaffold_1646</td>
<td>16 671</td>
<td>298–212</td>
<td><em>Ambidensovirus</em> Myzus persicae densovirus</td>
<td>NS</td>
<td>gil33235699</td>
<td>ref</td>
<td>NP_874375_1</td>
<td>51.7</td>
<td>29</td>
<td>8E–08</td>
</tr>
<tr>
<td>scaffold_1646</td>
<td>16 671</td>
<td>456–316</td>
<td><em>Ambidensovirus</em> Myzus persicae densovirus</td>
<td>NS</td>
<td>gil33235699</td>
<td>ref</td>
<td>NP_874375_1</td>
<td>34.7</td>
<td>49</td>
<td>8E–08</td>
</tr>
<tr>
<td>scaffold_1826</td>
<td>10 113</td>
<td>2042–1156</td>
<td><em>Ambidensovirus</em> Myzus persicae densovirus</td>
<td>NS</td>
<td>gil33235700</td>
<td>ref</td>
<td>NP_874376_1</td>
<td>37.5</td>
<td>664</td>
<td>5E–105</td>
</tr>
<tr>
<td>scaffold_1839</td>
<td>9 957</td>
<td>6335–7213</td>
<td><em>Ambidensovirus</em> Dysaphis plantaginea densovirus</td>
<td>NS</td>
<td>gil204324151</td>
<td>gb</td>
<td>ACI01073_1</td>
<td>63.0</td>
<td>292</td>
<td>2E–116</td>
</tr>
<tr>
<td>scaffold_1839</td>
<td>9 957</td>
<td>9111–7447</td>
<td><em>Ambidensovirus</em> Myzus persicae densovirus</td>
<td>CP</td>
<td>gil33235702</td>
<td>ref</td>
<td>NP_874377_1</td>
<td>40.6</td>
<td>571</td>
<td>7E–105</td>
</tr>
<tr>
<td>scaffold_1839</td>
<td>9 957</td>
<td>9505–9041</td>
<td><em>Ambidensovirus</em> Myzus persicae densovirus</td>
<td>CP</td>
<td>gil33235701</td>
<td>ref</td>
<td>NP_874378_1</td>
<td>33.7</td>
<td>166</td>
<td>2E–13</td>
</tr>
<tr>
<td>scaffold_1885</td>
<td>8 960</td>
<td>8151–8591</td>
<td><em>Ambidensovirus</em> Dysaphis plantaginea densovirus</td>
<td>CP</td>
<td>gil204324154</td>
<td>gb</td>
<td>ACI01076_1</td>
<td>51.0</td>
<td>149</td>
<td>4E–38</td>
</tr>
</tbody>
</table>
reference sequences of MpDNV, DplDNV, JcDNV, GmDNV and HaDNV proteins (Table S2).

The probability that these endogenous viral sequences correspond to DLSs was reinforced by the prediction of functional ORFs corresponding to either CP or NS proteins (Fig. 2, Table S3). On scaffold_1839, three ORFs potentially encoded one NS and two CP proteins, whereas on scaffold_794 each densovirus full-length integration could encode two NS and one CP protein (Fig. 2, Table S3). DLS integrations were confirmed experimentally by PCR amplification on an M. persicae clone originating from Colmar (France). In order to avoid potential non-specific detection of non-integrated MpDNV sequences, free MpDNV was eliminated from this clone by isolating individual larvae immediately after birth and depositing them on sugar beets (see Methods). A first set of primers was designed to specifically amplify the internal integrated sequences showing the highest nucleotide identity with the aphid DNV sequences (primers referred to as ‘integration’ in Table S4; primer positions shown in Fig. 2a). As shown in Fig. 2(b, upper panel), the presence of the integrated viral sequences was confirmed by amplification from the virus-free aphid genome of DNA fragments of the expected sizes (Table S4). The pair of primers for PXV amplified two fragments on the M. persicae genome: one of the expected size (142 bp) and a second representing a longer fragment (~200 bp), probably corresponding either to a non-specific hybridization in the aphid genome (Fig. 2b, upper panel) or to a larger sequence that is fragmented in the current genome assembly upon which fragment size predictions were based. A second set of primers was designed to amplify DNA fragments encompassing insect and viral-like sequences (primers referred to as ‘junction’ in Table S4; primer positions shown in Fig. 2a). These fragments of larger size (1054–4131 bp) were also amplified from the M. persicae genome, although with a lower efficiency when using the two pairs of primers for P3 and PIX (Fig. 2b, lower panel). Taken together, these results confirmed the presence of DLSs in the genome of M. persicae.

**Phylogenetic analysis**

To evaluate the genetic relationships between DLSs present in both the M. persicae and A. pisum genomes with known DNV sequences, we performed a phylogenetic analysis using all integrated sequences that contained a 300 bp DNA sequence from the NS1 gene. This sequence corresponds to highly conserved domains (DNA helicase and ATPase domains) of the rep gene in parvoviruses. As the nucleotide sequences were too divergent to be reliably aligned, we used the translated sequence of all 300 bp sequences (100 aa) to build a phylogenetic tree. Two nodes potentially corresponding to different integration events were observed (Fig. 3). The first cluster contained a mixture of DLSs found in the M. persicae and A. pisum genomes that grouped with MpDNV and DplDNV (Hub 1, bootstrap value 93%). In particular, M. persicae DLSs seemed to be closely related to these two aphid DNVs (bootstrap value 82%). The second cluster was almost exclusively composed of DLSs from A. pisum, and might result from a different integration event (Fig. 3). Only one DLS in the M. persicae genome (MpDLS_28) grouped in this cluster, but the tree robustness at this level (Hub 2, bootstrap value 15%) was not sufficient to confirm with confidence the affiliation of MpDLS_28 to this cluster.

**Transposable element (TE) detection in proximity to DLS integrations**

We addressed the question of the frequency of TEs in proximity to DLS integrations to identify elements possibly responsible for the integration of the infectious virus sequence into the host genome (endogenization). Some paroviruses appear to have evolved towards a transposon-like propagation during evolution, as was shown in
Fig. 2. (a) Schematic representation of DLSs in scaffold_1839 (upper panel) and scaffold_794 (lower panel) of the *M. persicae* genome (clone G006). Shaded arrows on the black line stand for potential ORFs of DLS (NSint and CPint). The percentage identity with the corresponding sequence in the genome of MpDNV is indicated above. The nucleotide position of these potential ORFs in the scaffold together with the first methionine ([INSERT Fig. 002] [INSERT Fig. 001]) in the sequence is indicated. The MpDNV genome organization is presented below with variously shaded arrowheads. The position of primers used to confirm DLS integrations and transcription is shown by short arrows. (b) Confirmation of DLS integration in scaffold_1839 and scaffold_794 in the genome of *M. persicae* DNA (clone Colmar) by PCR amplification of internal integrated viral sequences (upper panel) or junctions between DLSs and the *M. persicae* sequence (lower panel). PCR products were separated by gel electrophoresis on a 2 % (upper panel) or 0.7 % (lower panel) agarose gel. The names of primers and the targeted scaffolds (S) are indicated above each lane. The position of molecular mass markers (lane L) is shown on the left. The asterisk indicates a non-specific amplification, while arrowheads show DNA fragments of low intensity. The primer sequences are listed in Table S4.
We investigated the presence of TEs in genomic sequences of the clone G006 of *M. persicae* covering 30 kb upstream and downstream of each DLS using Repbase and National Center for Biotechnology Information (NCBI) sequence analysis tools (BLASTX and BLASTN) (Table S5). In four scaffolds (scaffold_794, _413, _1335 and _883), we detected TEs on both sides of DLS integrations (Fig. 1). On the other scaffolds, only one TE was located at either the 3' or 5' side of the DLS (Fig. 1). Several full-length (or almost-complete) sequences of TEs were present in some scaffolds (Table S5). Most of the putative TEs were DNA transposons from class II in the classification originally proposed by Finnegan (1989). Despite the presence of TEs in scaffolds containing DLSs, their presence could not be correlated with the presence of DLSs, as a similar frequency of TE integrations was also observed in DLS-free regions (data not shown). We also did not observe terminal inverted repeats flanking the DLSs, which are hallmarks of proliferation by transposition (data not shown).

**Fig. 3.** Phylogenetic relationships of DLS integrations in *M. persicae*, *A. pisum* and exogenous DNVs (underlined). DLS nomenclature is consistent with Tables 1 and S2. Bootstrap values correspond to 1000 replications and are indicated for values >50%. Sequences of 100 aa corresponding to translation of the integrated DLSs in *M. persicae* (see Table 1) and *A. pisum* genomes (see Table S2) sharing similarity with NS1 of parvoviruses were aligned with densovirus sequences arising from: *Aedes albopictus* densovirus (*AaDNV*; GenBank accession no. NC_004285.1), infectious hypodermal and haemato-poietic necrosis virus (*IHHNV*; NC_002190.2), *Bombyx mori* densovirus 5 (*BmDNV5*; NC_004287.1), *Fenneropenaeus chinensis* hepatopancreatic densovirus (*FchDV-1*; NC_014357.1), *JcDNV* (NC_004284.1), *MpDNV* (NC_005040.1), *DpIDNV* (EU851411.1), *PfDNV* (NC_000936.1), *GmDNV* (NC_004286.1) and *HaDNV* (JQ894784.1). *Chicken parvovirus* NS1 (AJB28748.1) was used to root the tree. The scale represents probabilities of change from one amino acid to another in terms of a unit, which is an expected 1% change between two amino acid sequences.
Transcription of DLSs in *M. persicae*

We addressed whether DLSs in the *M. persicae* genome were transcribed by performing reverse transcription (RT)-PCR and real-time RT-PCR experiments on polyadenylated mRNA extracted from whole aphids (clone Colmar deprived from MpDNV) (see Fig. 2 for position of the potential DLS ORFs). Absence of residual genomic DNA was first confirmed by a lack of amplification when the PCRs were performed directly on total RNA (data not shown). Primers (Table S4) were designed to specifically analyse transcription of each of the potential viral coding sequences on scaffold_1839 and _794. The expected DNA fragments were obtained for the NSint_1 and CPint_2 ORFs on scaffold_1839 or from the CPint_1 and CPint_2 ORFs on scaffold_794 (data not shown). Amplification was also observed using the primer pair for Py (Fig. 2), which suggested a potential transcription of NSint_1 and/or NSint_4. Indeed, the high sequence identity between these two ORFs prevented detection of each specific mRNA. The specificity of the amplification reactions was confirmed by sequencing the amplified fragments. In contrast, no fragments or fragments of unexpected size were obtained using the other sets of primers amplifying potential transcripts from CPint_1 on scaffold_1839 or NSint_2 and/or NSint_3 on scaffold_794, suggesting an absence of transcription of these potential ORFs (data not shown). In a second step, in order to measure the rate of accumulation of the viral transcripts detected by RT-PCR, four independent real-time RT-PCRs were performed on polyadenylated mRNAs extracted from whole aphids, and the results of one representative experiment is presented in Fig. 4. The relative accumulation of mRNA transcribed from DLS ORFs was normalized to two housekeeping genes (RLP7 and L27). Interestingly, we observed amplification of mRNA for each of the DLS integrations assayed (Fig. 4). These expressions were confirmed in three additional experiments, although some variation in the transcription levels of CPint_1 on scaffold_794 and of CPint_2 on scaffold_1839 was observed between the different biological replicates (Fig. S1). Despite these variations, the tendency was in favour of a higher transcription of the DNV CP-like genes compared with the DNV NS-like genes.

**DISCUSSION**

Our study provides evidence of the presence of DLSs in the *M. persicae* genome. Integration of viral genomes belonging to the subfamily Densovirinae (family Parvoviridae) into arthropods has been reported previously and, in particular, DLSs have been identified in the genome of the pea aphid, *A. pisum* (Liu et al., 2011b). The presence of DLSs in the genome of *M. persicae* implies that the ancestor of *M. persicae* was infected with a DNV whose genome was endogenized. Three almost-complete ambidensovirus-like genomes were integrated into the genome of the two sequenced *M. persicae* clones, which have distinct geographical origins (clone O from Europe and clone G006 from the USA). This suggests old and stable integration events. The high capacity of paroviruses to integrate into their host genome is probably due to the virus replication cycle in the nucleus, which requires synthesis of a double-stranded replicative form and the presence of inverted terminal repeats that facilitate integration into the host DNA (Berns, 1990; Yang et al., 1997). The mechanism by which DLS integrations occurred in the *M. persicae* genome is still unknown but is not thought to be linked to TEs, as no hallmarks of transposition were found in the proximities of DLSs. It is therefore likely that the integration paradigm of adeno-associated virus, which has been extensively studied, could apply to other members of the family Parvoviridae including DNVs. Adeno-associated virus integration relies on the endonuclease activity of its rolling-circle replication protein, a function that is shared with the NS1 protein of DNV (Ding et al., 2002; Krupovic & Forterre, 2015).

Interestingly, potential coding capacities retained within the DLSs were identified, and transcription in *M. persicae* of the integrated NS- or CP-encoding sequences (hereafter referred to as NSint and CPint) was confirmed. As the predicted viral ORFs were not all transcribed, and the transcript levels from transcribed ORFs appeared unequal,
the presence of functional promoters was assessed (Neural Network Promoter Prediction software, http://www.fruitfly.org/seq_tools/promoter.html; data not shown). In scaffold_1839, potential promoters upstream of CPint_2 were identified, whereas no promoter sequences were detected upstream of CPint_1. This observation could explain the lack of CPint_1 transcription observed in our experiments. On the same scaffold, potential promoters that could control transcription of NSint_1 were also found. On the other scaffold (scaffold_794), promoter-like sequences with the potential to control transcription of CPint_1 and CPint_2 were found upstream of each ORF. Transcription of NSint_1 and/or NSint_4 could also be explained by the presence of promoter-like sequences upstream of these two ORFs. As occurs in MpDNV, NS ORFs are supposed to be transcribed from a single promoter (van Munster et al., 2003b), so transcription of the downstream NSint_2 and NSint_3 was expected but not observed in our experiments. The absence of transcription of these ORFs could indicate a transcription termination signal between the two NSint ORFs. Altogether, promoter sequence identification coincided with our detection of viral transcripts. All potential promoter sequences driving transcription of the endogenous viral ORFs are probably of aphid origin, as they do not share any sequence similarity with the predicted promoters in MpDNV or with other viral sequences of MpDNV or DplDNV (data not shown). The higher level of transcription of CP-like genes compared with NS-like genes could therefore reflect a difference in endogenous host promoter transcription efficiency. Transcription of endogenized parovirus sequences in A. pisum was also suggested previously, as expressed sequence tags exhibiting similarities to parovirus-related DNA were reported in the NCBI public databases (Liu et al., 2011b). However, no experimental confirmation of mRNA accumulation of these parovirus-related DNAs was reported. Although we clearly detected polyadenylated RNA originating from DLSs in M. persicae, the question of their translation remains.

Transcription, and potentially translation, of these sequences could explain the low pathogenicity of DNV in aphids. To date, there are only two DNVs described in aphids: MpDNV, which is highly specific to M. persicae (van Munster et al., 2003a), and DplDNV, identified in D. plantaginea (Ryabov et al., 2009). Interestingly, these aphid DNVs do not harm their host, in contrast to other DNVs which are highly pathogenic to their insect hosts (Barreau et al., 1996; Liu et al., 2011c; Mutuel et al., 2010; Suto, 1979). MpDNV and DplDNV only slightly affect the development or reproduction of aphids (Ryabov et al., 2009; van Munster et al., 2003a). Moreover, the existence of a selective advantage for the aphid to harbour the virus has been shown for DplDNV, which confers a beneficial effect on D. plantaginea by inducing the production of winged morphs at a low insect density, therefore promoting aphid dispersal (Ryabov et al., 2009). Concerning MpDNV, although there is a clear advantage for the virus to maintain a suboptimal infection that does not kill the host (van Munster et al., 2003a), no beneficial effect of MpDNV on M. persicae has been discovered so far (V. Brault, unpublished).

The cause for latency in DNV infections of these viruses within aphid hosts could be due to DLS integrations in the aphid genome, as reported in this study. Indeed, transcription of these viral integrated sequences could prime a silencing-based defence mechanism that could target cognate viruses and inhibit the development of an acute infection, as has been shown for Drosophila (Goic et al., 2013) and bees (Maori et al., 2007). In this respect, DLSs in scaffold_794 or _1839 contained 12 stretches of 21–29 nt sharing either a perfect nucleotide sequence identity with MpDNV or having only one or two mismatches with MpDNV or DplDNV (Table S6). Although these sequence homologies were limited to a small portion of the viral genome, they were of an appropriate length to induce viral RNA degradation, or translation inhibition, which could control and limit infectious DNV multiplication. Another mechanism of protection induced by the integrated viral sequences could be the synthesis of non-functional DNV-like proteins that could inhibit virus multiplication, as has been shown for human immunodeficiency virus type 1 (HIV-1). Expression of a Gag protein from a human endogenous retrovirus impaired progeny release and HIV-1 infectivity (Monde et al., 2012). The formation of empty particles from endogenized CP protein genes able to prime insect immune protection against further infections could also be an alternative hypothesis explaining protection conferred by viral endogenized sequences. In addition, the activation of the immune system of the insect could be responsible for reducing the pathogenicity of the DNV. These different aspects of virus control should be addressed in the future.

METHODS

Insect rearing and virus clearing. An M. persicae (Sulzer) clone from Colmar (France) was reared on pepper (Capsicum annuum) at 20 °C with a 16 h photoperiod. This clone was infected with MpDNV. The same clone was cured of MpDNV by transferring individual larvae directly after birth onto beets (Beta vulgaris) to avoid contamination by contact. The absence of MpDNV was monitored by PCR (van Munster et al., 2003a), and one aphid colony cured of MpDNV was maintained on B. vulgaris.

DLS detection in the M. persicae and A. pisum genomes. To screen for DLSs in the M. persicae genome, BLASTN and tBLASTn searches were performed online against the AphidBase database (http://www.aphidbase.com/aphidbase/), using M. persicae clone G006 (originating from the USA; Ramsey et al., 2007) as a reference genome (scaffolds). Full DNA genomes of the following subfamily Densovirinae member types: infectious hypodermal and haematopoietic necrosis virus (IHHNV, genus Pentastyledenovirus), Aedes albopictus densivirus (AaDNV, genus Brevidensovirus), Bombyx mori densivirus 5 (BmDNV5, genus Iteradensovirus), Fennneropenaeaeus chinensis hepadensovirus (FchDV-1, genus Hepadensovirus) and
JcDNV (genus Ambidensiovirus), as well as full DNA genomes and NS and CP protein sequences of MpDNV and DplDNV, were used as queries. All non-redundant hits from these searches with E-values of $<1.10^{-3}$ were extracted and used in further analyses.

To eliminate possible chimeras or errors in DLS regions of the genome assembly, BLAST searches using extracted DLSs from *M. persicae* clone G006 were performed against the genome of *M. persicae* clone_O (available in AphidBase and originating from the UK). In a similar manner, BLAST searches were performed on the updated assembly of the *A. pisum* genome (assembly v.2) also available in AphidBase. *A. pisum* sequences showing similarity to subfamily Densovirinae representatives were also identified and extracted for phylogenetic analysis.

Integration of DLSs in *M. persicae* clones was confirmed by amplifying virus-like sequences using specific primer pairs designed with the Primer3 program (Rozen & Skaletsky, 2000) and with binding sites in both flanking host and virus insertion sequence (Table S4, Fig. 2). Total genomic DNA was extracted from 30 aphids, which were ground in liquid nitrogen using the QIAamp DNA Mini kit protocol (Qiagen). One microlitre of DNA (20–30 ng µl$^{-1}$) was used for PCRs containing the specific primers and performed in a thermal cycler (Eppendorf) using standard protocols. PCR products were analysed by 0.7 or 2% agarose gel electrophoresis depending on amplicon size and visualized under UV after ethidium bromide staining.

**Alignment and phylogenetic reconstruction.** Phylogenetic trees were reconstructed using the predicted amino acid sequences of DLSs. Protein sequences of 100 aa corresponding to a region of the DLS sharing similarity with NS1 of parvoviruses were aligned with the corresponding sequences of known DNVs including subtypes of the subfamily Densovirinae using the MUSCLE program configured for high accuracy (Edgar, 2004). The DNV sequences used in the alignment were: AdaDNV (GenBank accession no. NC_004285.1), HHINV (NC_000190.2), BmDNV5 (NC_004287.1), FchDV-1 (NC_014357.1), JcDNV, (NC_004284.1), MpDNV (NC_005404.1), DplDNV (EU851411.1), Periplaneta fuliginosa densovirus (NC_00936.1), GmDNV (NC_004286.1) and HaDNV (JQ894784.1). Phylogenetic relationships among DLS integrations and DNVs were assessed using the Phylogeny.fr platform (Dereeper et al., 2008). The phylogenetic tree was reconstructed using the maximum-likelihood method implemented in the PhyML program (v.3.1/3.0 aLRT). The JTT 1 G model (Jones et al., 1992), a commonly used probabilistic amino acid replacement matrix at the 85% identity level allowing heterogeneity across sites, was selected. Reliability for the internal branch was assessed using the aLRT test (SH-Like). Node support was estimated using bootstrap resampling procedures (1000 trials). Trees were edited using TREEDyn (v.198.3) (Chevenet et al., 2006).

**Detection of TEs in the *M. persicae* genome.** A 30 000 bp region encompassing each DLS was scanned for adjacent TEs or repetitive sequences using CENSOR (Repbase; http://www.girinst.org/censor/index.php) with a reference collection of sequence repeats (Kohany et al., 2006). BLASTX and tBLASTx queries on these flanking DLS regions were also performed. As a control, repetitive sequences in DLS-free regions of scaffolds harbouring DLSs and on scaffolds where no DLSs could be detected were also analysed.

**Transcription of DLSs in *M. persicae*.** Transcription of DLSs was analysed by real-time RT-PCR in fourth-instar individuals of *M. persicae* (clone Colmar). Total RNA was extracted from aphids cured of MpDNV (four biological replicates) ground in liquid nitrogen following the Qiagen RNeasy Plant Mini kit protocol (animal tissue protocol). On-column DNase treatment was carried out using an RNase-Free DNa se kit (Promega). Synthesis of cDNA (from 2 µg RNA) was primed using oligo(dt) (Promega) and cDNA synthesis was done using the Moloney murine leukemia virus reverse transcriptase (Promega). The cDNA was then used as a template for real-time PCR amplification (primers listed in Table S4) using an iTaq universal SYBR Green supermix (Bio-Rad Laboratories). Ribosomal proteins RLP7 and L27 were used as internal controls. The amplification reaction consisted of 10 × SYBR Green supermix and 10 pmol forward and reverse primers (Table S4) in a final volume of 20 µl. Three technical repeats were performed using a CFX96 Touch Real-Time PCR Detection System (Bio-Rad Laboratories). The mRNA synthesis was normalized with RLP7 and L27 using the 2$^{-ΔΔCt}$ method (Livak & Schmittgen, 2001) and analysed with Bio-RadCFX Manager software (Bio-Rad Laboratories).

**ACKNOWLEDGEMENTS**

DNA sequence data were downloaded from AphidBase (http://www.aphidbase.com/aphidbase/). Funding for *M. persicae* clone G006 genomic sequencing was provided by USDA-NIFA award 2010-65105-20558 held by Alex Wilson (Miami University, OH, USA). Funding for *M. persicae* clone_O genomic sequencing was provided by The Genome Analyses Centre (TGAC) Capacity and Capability Challenge program project CCC-15 held by Saskia Hogenhout and from the Biotechnology and Biological Sciences Research Council (BBSRC) (BB/004553/1) and the John Innes Foundation. We are grateful to Fabrice Legeai and Antony Bretandeau (INRA, Rennes) for their assistance in aphid genomic sequences analysis. We also thank Denis Filloux (INRA, Montpellier, France) for his help in database analysis, Neta Luria and Aviv Dombrovsky (Volcani Center, Israel) and Pierre Gladieux (INRA, Montpellier, France) for their fruitful discussions. We are grateful to Keith Garrisson (Saint Mary’s College of California, CA, USA) and A. Wilson for critically reading this manuscript.

**REFERENCES**


