Unusual genomic features of a badnavirus infecting mulberry

Michela Chiumenti, Massimiliano Morelli, Angelo De Stradis, Toufic Elbeaino, Livia Stavolone, and Angelantonio Minafra

INTRODUCTION

Badnaviruses are plant-infecting pararetroviruses belonging to the family Caulimoviridae consisting of circular, dsDNA genomes varying from 7.0 to 9.2 kb. Each strand of the genome is not covalently closed but contains a single site-specific discontinuity (Medberry et al., 1990). Genomic DNA is contained in non-enveloped bacilliform particles 30 nm in width, with a modal length of 130 nm, albeit virions ranging in length between 60 and 900 nm were described (Geijskes et al., 2004). Most of the badnavirus genomes carry three main ORFs on the positive strand (Hull, 2001): ORF1, coding for an unknown protein with a conserved domain (DUF1319; Sether et al., 2012); ORF2, the virion-associated protein (VAP); and the longest, ORF3, encoding a polyprotein hosting several functional and structural domains in the following order: movement protein (MP), coat protein (CP), aspartic protease, retrotranscriptase (RT) and RNase H (Hohn & Rothnie, 2013; Hohn et al., 1997). Some badnaviruses encode one or more smaller ORFs of unknown function (Kazmi et al., 2015). Having penetrated the host cell, pararetroviruses replicate via a mechanism of transcription/reverse transcription like mammalian retroviruses, albeit not going through an integration step in the host genomes. The dsDNA genome is released from the capsid, the discontinuities are repaired by repair polymerases and ligase and, finally, DNA is transferred into the nucleus. The circular minichromosome produced in this way is the template for the transcription of the greater-than-length pregenomic/polycistrionic mRNA. The pregenomic RNA (pgRNA) moves within the cytoplasm where it serves either as a template for either the synthesis of new genomic molecules or the translation of viral proteins (Hohn & Rothnie, 2013; Hohn et al., 1997).

Badnaviruses are both triggers and targets of RNA silencing (Blevins et al., 2006; Kazmi et al., 2015), the RNA-based and sequence-specific mechanism of gene regulation reported for most eukaryotes (Chen, 2009; Voinnet, 2008). Gene silencing can occur at two levels: (i) it can inhibit transcription processes (transcriptional gene silencing) by DNA methylation and/or histone modifications of the DNA targets or (ii) at the post-transcriptional level (post-transcriptional gene silencing), causing inactivation or degradation of target transcripts (Bologna & Voinnet, 2014; Parent et al., 2012; Pumplin & Voinnet, 2013). The specificity of this mechanism is controlled by small RNAs (sRNAs): short RNA fragments 21–24 nt long, which in plants are generated by Dicer-like proteins (DCL) cleaving dsRNAs or highly structured ssRNAs. sRNAs are loaded in Argonaute...
(AGO) proteins for targeting partially or wholly complementary nucleic acids for post-transcriptional or transcriptional silencing (Bologna & Voinnet, 2014; Mallory & Vaucheret, 2010; Parent et al., 2012). The application of next-generation sequencing (NGS) technology makes it possible to characterize virus-derived sRNA and reconstruct complete virus genomes (even unknown ones) by alignment and assembly of sequenced viral sRNAs (Adams et al., 2009; Barba et al., 2014; Massart et al., 2014). Blevins et al. (2006, 2011) thoroughly investigated pararetrovirus sRNA accumulation in Arabidopsis thaliana lines carrying mutations for some of the enzymes involved in RNA silencing pathways, and indicated the increased amount of 24-mer sRNAs as the specific landmark of nuclear replicating viruses, involved in RNA-directed DNA methylation.

NGS also allows more detailed investigation of the heterogeneity of subgenomic nucleic acids in enriched preparations from virus-infected plants (Chiumenti et al., 2016a; Ng et al., 2011). Subgenomic RNAs and defective RNAs or DNAs can be distinguished in four classes: (i) satellite viruses, encoding a structural protein specifically encapsidating their genomes and producing nucleoprotein components distinct from those of the helper viruses; (ii) satellite RNAs, encoding non-structural proteins or no proteins at all, and encapsidated by the CP of the helper viruses; both satellite viruses and RNAs depend on their helper virus for replication, but they also usually share little sequence similarity with them apart from the terminal regions (Simon et al., 2004); (iii) satellite DNAs, associated only with Geminiviridae (Rojas et al., 2005) and encoding a functional protein necessary for the biological success of the associated virus. These can be distinguished into alpha- and beta-satellites; the former encoding their own replicase protein, are able to autonomously replicate in the host but require the helper virus for spread in plant and insect transmission; the latter, instead, depend on the begomovirus for their replication, while the helper virus needs the beta-satellite for efficient host infection; (iv) defective interfering (DI) and defective (D) RNAs, directly deriving from their 'helper' virus genomes, being produced by replication errors of their parent (helper virus); these usually do not depend on their helper virus for replication, but the parent virus provides the missing replication proteins in trans (Pathak & Nagy, 2009; Simon et al., 2004).

In this study, we describe the biological and genomic features of a new badnavirus, Mulberry badnavirus 1 (MBV1), recently identified in a Morus alba L. plant in Lebanon, showing symptoms of leaf mottling and vein yellowing (Elbeaino et al., 2013). Sequencing of the complete virus genome, through an sRNA high-throughput sequencing (HTS) approach, revealed the existence of a full-length and a deleted form of the viral genome. Either the full-length or shorter MBV1 genomes can be independently encapsidated into virus particles, and both genome forms are infectious and systemically translocate in infected mulberry plants.

**RESULTS**

**MBV1 genome-sequencing strategy**

We recently reported mulberry disease associated with the presence of the tentative new virus species MBV1. With the purpose of characterizing the MBV1 genome and understanding its interaction with the host plant, an sRNA library was sequenced from the MBV1-infected mulberry L34. A total of 4,513,704 raw short reads were obtained. Quality filter steps and adaptor removal resulted in 4,321,760 redundant reads, corresponding to 1,193,828 non-redundant (nr) reads. As expected, the library profile showed the bimodal distribution, with a relative abundance of peaks at 21, 22 and 24 nt, either as redundant or nr reads.

All quality-filtered reads were subjected to de novo assembly using a 17k-mer length to generate 1977 contigs, ranging from 33 to 175 nt (average size 58,027 nt). Blast analysis identified 77 contigs out of the total, 4 of which, in a range of 36 to 61 nt, hit for a viral origin and specifically recognized badnavirus sequences. The remaining contigs showed a plant origin.

We used the complete genomes of CSSV, CiYMV and FBV1 as reference sequences to perform a guided assembly with the SOAP aligner program (Li et al., 2008). We found a total of 4,521 redundant reads, corresponding to 0.10 % of the total library, matching on the CSSV genome: 0.13 % aligned on the CiYMV genome and 0.26 % to FBV1. All reads showed a hotspot organization on each of the reference genomes considered. As a greater number of reads recognized FBV1, we selected the hotspot reads aligning to this virus genome for primer design.

**MBV1 genomic features**

Starting with primers designed on reads matching the FBV1 sequence, we performed genome walking sequencing to reconstruct the complete genome sequence of MBV1.

The sequence assembly resulted in 6,945 nucleotides in length with a G+C content of 44.49 %. The viral genome, deposited in the EMBL database with the accession number LN65125, encodes a single polyprotein between nt 404 and nt 6403 with a predicted molecular weight of 229.99 kDa (Fig. 1a). A second hypothetical ORF was found in positions 6088 to 6441 (TAG stop codon), with reading frame +1, encoding an unknown protein of 119 aa and a predicted molecular weight of 13.62 kDa. No ORF-encoding proteins larger than 10 KDa were found in the minus strand.

The complete genomic sequence contains the negative-strand tRNA\textit{MET}-binding site (TGGTATCAGAGC\textit{TCGG}), priming minus-strand synthesis during replication of the genome and is conventionally considered as the starting point for the genome sequence (Fig. 1a; Bouhida et al., 1993). The intergenic region downstream of the polyprotein hosts a putative TATA box (TATAAG\textit{AAGA}), an essential part of the badnavirus promoter region (Medberry et al.,
Fig. 1. Relevant genomic features of MBV1. (a) Graphical representation of the genome organizations of members of the Caulimoviridae family. Boxes represent ORFs, with respective numbers indicated at the top. The position of conserved motifs (described at the bottom) is shown on each genome. Accession numbers and genome length are indicated for each virus on the right. Grey numbered arrows below MBV1 genome indicate the position of predicted coiled coils. Light blue arrows represent the primers flanking the gap region in the deleted genome. Blue and light blue rectangles indicate the different amplicons obtained with the same couple of primers (Gap-for/Gap-rev) from the two genomic templates. Amplicon dimensions are |

### Table: Conserved Motifs

<table>
<thead>
<tr>
<th>Motif</th>
<th>Start</th>
<th>Stop</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motif 1</td>
<td>100</td>
<td>116</td>
<td>IPIAEIQIYRRTNARR</td>
</tr>
<tr>
<td>Motif 2</td>
<td>396</td>
<td>416</td>
<td>ERPSQDLADKSYAVETI</td>
</tr>
<tr>
<td>Motif 3</td>
<td>959</td>
<td>978</td>
<td>SISNEDYQNEQLLISYVF</td>
</tr>
<tr>
<td>Motif 4</td>
<td>1350</td>
<td>1368</td>
<td>TVAAikkREvNYKELQDI</td>
</tr>
<tr>
<td>Motif 5</td>
<td>1705</td>
<td>1722</td>
<td>NKBVHELKAMDLKIN</td>
</tr>
</tbody>
</table>

http://jgv.microbiologyresearch.org
indicated. (b) Amino acid alignment of biologically significant homologous regions of proteins encoded by representative members of the Caulimoviridae family. Before each sequence the virus acronyms are indicated, as well as the ORF designation and the number of residues separating the protein N-terminus from the aligned segment. Boxes denote identical amino acids, also shown in italics below the alignment. The spacing between amino acid blocks is given in angle brackets. (c) Prediction of coiled-coil domains in the MBV1 ORF1. Coiled-coil motifs predicted in MBV1 ORF1 are aligned in the last column. The seven positions of the heptad repeat are labelled a–g in the bottom row of the alignment. Positions a and d, occupied by hydrophobic residues that are part of the oligomer core, are highlighted; the other positions are predominantly solvent-exposed polar residues. First and last amino acid numbers are indicated in columns 3 and 4, respectively. (d) Representation of the long intergenic region (LIGR). The region is comprised between 31 nt downstream of the TATA box sequence and the beginning of ORF1. Short ORFs are depicted by small blue empty boxes. Predicted sORFA is represented by a red box. The thick black arrow represents the tRNA<sub>MET</sub>-binding primer, while dotted lines on the drawing define the complementary sequences that form the base of the large hairpin structures. mFold prediction of the large stem-loop structure is represented in a dotted box with its free energy value indicated above. The scale bar underneath the LIGR gives the dimension of the leader sequence.

1992). Even if the exact promoter location needs to be confirmed experimentally, other conserved elements of the putative promoter were identified, such as the nucleotide motifs CACAAT (5734–5739) and TGACG (4899–4903), upstream of the TATA box (Odell et al., 1985) and the putative polyadenylation sequence localized downstream of the promoter (AATAAA<sub>863–869</sub>) (Geering et al., 2000).

Analysis of the single polyprotein revealed the presence of all the key badnavirus-conserved motifs: a portion of the DUF1319 (domain of unknown function distinctive of the badnaviruses) found in an unconventional position, indicated in italics below the alignment. The spacing between amino acid blocks is given in angle brackets. (c) Prediction of coiled-coil domains in the MBV1 ORF1. Coiled-coil motifs predicted in MBV1 ORF1 are aligned in the last column. The seven positions of the heptad repeat are labelled a–g in the bottom row of the alignment. Positions a and d, occupied by hydrophobic residues that are part of the oligomer core, are highlighted; the other positions are predominantly solvent-exposed polar residues. First and last amino acid numbers are indicated in columns 3 and 4, respectively. (d) Representation of the long intergenic region (LIGR). The region is comprised between 31 nt downstream of the TATA box sequence and the beginning of ORF1. Short ORFs are depicted by small blue empty boxes. Predicted sORFA is represented by a red box. The thick black arrow represents the tRNA<sub>MET</sub>-binding primer, while dotted lines on the drawing define the complementary sequences that form the base of the large hairpin structures. mFold prediction of the large stem-loop structure is represented in a dotted box with its free energy value indicated above. The scale bar underneath the LIGR gives the dimension of the leader sequence.

Caulimoviridae in the RT/RNase H region (4564–5586 nt). The obtained cladogram denotes that MBV1 locates in a branch comprising CSSV, CiMV and FBV1 (Fig. 2). Nucleotide pairwise comparison with the full-length sequences of the badnaviruses included in the same clade revealed that MBV1 has 58.1 % nucleotide identity with CSSV, 57.4 % with CiYMV, 55.6 % with ComYMV and the highest identity with FBV1 (58.3 %). BLASTN analysis of the RT/RNase H region demonstrated that MBV1 shares less than 80 % nucleotide identity with any close badnavirus, which is the species demarcation threshold established by the ICTV guidelines (King et al., 2011).

Analysis of the virus-derived sRNAs in MBV1-infected mulberry

Based on the complete genome sequence of MBV1, we performed a detailed analysis of the distribution and characteristics of the virus-derived sRNA sequences.

A SOAP alignment of all the reads against the assembled reference sequence was computed, with two mismatches. Considering the circular conformation of the badnavirus genome, and in order to detect sRNA mapping over or between the conventional 5’ and 3’ of the genome, the first 24 nts of the 5’ termini were added at the 3’ end of the linear MBV1 full-length sequence. The genome was covered by 110 379 nr reads, equivalent to 1 049 717 redundant reads, scattered in hotspots along the entire genome sequence (Fig. 3a–c). A detailed characterization of the 20 to 24 nts revealed that 87 222 nr reads, corresponding to 945 053 redundant reads, aligned in this range size with a slightly higher accumulation of (+) (51.98 %) than (−) (39.04 %) MBV1 sRNA reads, either as nr or redundant (Fig. 3b, c).

Size class distribution of MBV1-sRNAs revealed prevalent peaks of 21 and 22 nt species (33.18 % and 26.32 % of total nr MBV1-sRNAs, respectively), but surprisingly, only 3.24 % of the total nr was represented by the 24-mer class (Fig. 3b), which is the hallmark of nuclear replicating viruses. In a parallel study, we obtained the sRNA profile of fig (accession <i>f5p5</i>) infected with FBV1 (Chiumenti et al., unpublished). Comparison of normalized size class distribution of sRNA
mapping reads from this latter virus to MBV1 confirmed the striking paucity of 24-mers in MBV1 even in relation to the phylogenetically closest badnavirus (Fig. 4).

Virus-derived sRNAs (vsRNA) show a preference for certain hotspots on the MBV1 genome (Fig. 3a). Differential vsRNA stability also contributes to hotspot formation. However, mapping the MBV1 sRNAs of both polarities revealed extensive targeting of the assembled virus genome, which was completely covered by overlapping nr sRNAs (Fig. 3a). In fact, the MBV1 sRNA profile revealed the presence of 21, 22 and 24 nt peaks overlapping each other (Fig. 3a), indicating that different DCLs, responsible for the biogenesis of each size class, may preferentially target the same specific regions of MBV1-derived dsRNA, influenced by template nucleotide composition and/or structure.

The variability of the 5′ nt, together with the length, are important hallmarks for the sorting of sRNAs into the AGO complexes (Kim, 2008; Mi et al., 2008). Therefore, a study of the 5′ nt of the MBV1 sRNA in the size class among 21 and 24 nt was done, evaluating (+) and (−) strands separately (Fig. 3d). Based on the sequence polarity, different trends were detected in (−)- and (+)-stranded sRNAs. In (−)-strand sRNAs, independently of the size class, the U nucleotide was prevalent, followed by C, A and G, whereas in the (+)-stranded sRNAs, 21-mers showed a majority of C, followed by A, U and G. The majority of 24-mers had an A, followed by C, U and G as 5′ nt, whereas the most frequent nucleotide for 22-mers and 23-mers was C, followed by A, U and G (Fig. 3d). MBV1 sRNAs with G at the 5′ terminal nucleotide were the least represented in each size class of both polarity strands (Fig. 3d).
Fig. 3. Analysis of MBV1-derived sRNAs. (a) Small interfering RNA (siRNA) mapping on linearized full-length MBV1 genome. Peaks result from the sum of every single nucleotide for each read. Differently coloured lines indicate different sizes among the

(a) Mulberry badnavirus 1 (LN651258)

(b) Redundant reads

(c) NR reads

(d) 5' Nucleotide study of MBV1 mapping reads

<table>
<thead>
<tr>
<th>Location on genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
<tr>
<td><strong>Abundance of reads</strong></td>
</tr>
<tr>
<td>10,000</td>
</tr>
<tr>
<td><strong>Location on genome</strong></td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td><strong>Abundance of reads</strong></td>
</tr>
<tr>
<td>250,000</td>
</tr>
<tr>
<td><strong>Location on genome</strong></td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td><strong>Abundance of reads</strong></td>
</tr>
<tr>
<td>15,000</td>
</tr>
<tr>
<td><strong>Location on genome</strong></td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td><strong>Abundance of reads</strong></td>
</tr>
<tr>
<td>80,000</td>
</tr>
<tr>
<td><strong>Location on genome</strong></td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td><strong>Abundance of reads</strong></td>
</tr>
<tr>
<td>80,000</td>
</tr>
<tr>
<td><strong>Location on genome</strong></td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td><strong>Abundance of reads</strong></td>
</tr>
<tr>
<td>80,000</td>
</tr>
<tr>
<td><strong>Location on genome</strong></td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td><strong>Abundance of reads</strong></td>
</tr>
<tr>
<td>80,000</td>
</tr>
</tbody>
</table>
Different sizes of MBV1 genome replicates in mulberry seedlings

MBV1 sequencing revealed an unconventional genomic organization (Fig. 1a). To confirm the presence of one long ORF1 and the small ORF2, several amplification experiments were performed on extracts from mulberry L34 with primers designed around the 5' region of the genome, where potential additional ORFs were expected to be located. Denaturant agents, such DMSO and betaine, were used to avoid any structural coiling in total DNA, and cDNA reverse transcribed from total RNA after DNase I digestion was also amplified. The final re-sequencing of the full-length genome, obtained by primer walking of a single PCR amplicon (6363 nt in length), confirmed the genome organization of MBV1 and the gap in the DUF1319 domain, but revealed the presence of an additional, shorter, viral sequence of pMBV-6k clone was strictly preserved in the genome, with the presence of 62 scattered point mutations, all silent with respect to the polyprotein sequence. Similar to the full-length genome sequence, it is organized within a single ORF (from nt 404 to nt 5347, reading frame of +2, 1647 aa). None of the conserved motifs described was affected by the deletion, with the exception of the amino acidic fragment CKCYAC (CXXCX,C), which was deleted in the cysteine-rich motif of the CP ZF domain (CXXCX,C).

The infectivity of the pMBV-6k clone was confirmed by mechanical inoculation of mulberry plants. Three M. alba and one M. nigra seedlings out of eight inoculated with the pMBV-6k clone exhibited mild vein yellowing and leafroll symptoms, 2 months after inoculation (Fig. 5a), and contained MBV1 DNA in both inoculated and systemic leaves as confirmed by amplification of two different functional domains of the MBV1 genome (RNase H and DUF1319, Fig. 5b).

From the individual p5-infected seedling, the MBV1 genome was fully re-sequenced through amplification of five different overlapping PCR fragments. The full-length sequence consisted of 5884 nts in length and showed 99.92 % similarity with the pMBV-6k clone and 98.89 % nucleotide identity with the MBV1 sequence from the mother plant L34. Four single point mutations were detected along the genome sequence. The genomic organization of the pMBV-6k clone was strictly preserved in the p5 isolate, with the gap starting and ending at exactly the same positions.

Electron microscopic (EM) observations confirmed the presence of bacilliform virus particles in systemic leaves of symptomatic and PCR-positive plantlets inoculated with the pMBV-6k clone (Fig. 5c). These particles appeared slightly different in shape from the wild-type ones, particularly for a 40 % shorter diameter size (~20 nm) and a more relaxed conformation of the capsid ultrastructure (Fig. 5c). It was possible to measure only 5 particles of 20 nm and 9 of 30 nm (in a total of 35 preparations), with a relative change in length from 120 to 320 for the former and 180 to 290 for the latter.

Genome size (~7 kbp) was consistently amplified from both M. alba and Morus nigra plants infected with pMBV-7k, confirming the identity with the cloned sequence.

Two MBV1 genomic variants coexist in infected plant

Full-length re-sequencing and EM observations demonstrated the ability of the shorter genome to infect mulberry and to encapsidate in viral particles, albeit of different width compared with those found in the original infected plant source, L34 (Elbeaino et al., 2013). To further investigate the occurrence of the two genomic forms found in infected plants, a primer set was specifically designed on the gap-flanking regions to amplify two differently sized amplicons depending on the presence/absence of the gap (Fig. 1a). To
evaluate the relative proportion of the two MBV1 genomic forms in L34, 10-fold serial dilutions of L34 tDNAex and encDNA in purified virus particles were assayed by PCR. Both amplicons were detected at up to $10^{-6}$ template dilution (10$^{-4}$ for encDNA), but the concentration of the full-length genome was higher either in tDNAex or in encDNA extracts (Fig. 6a). As expected, sequences from the larger fragment resulted in 1859 nt and mapped from position 1775 to 3633 of the MBV1 reference sequence. Consistently, the smaller amplicon resulted in a length of 806 nt, aligning from nt 1775 to 2083 and from position 3136 to 3633 of the MBV1 sequence. These results confirmed that both genomic variants accumulate in infected plants, and the ratio between them is also constant in encapsidated particles.

Southern blot hybridization, performed on L34 tDNAex and encDNA, revealed a specific, clear-cut reaction with the MBV1 genome. In particular, the virus-specific probes did not react on plant genomic DNA, either from healthy or L34 samples. L34 total DNA gave hybridization signals for two bands at about 7 and 6 kb, with the upper showing a more intense signal than the lower (Fig. 6b). The different signal intensity observed confirmed the ratio of the viral genomic forms observed by PCR. DNA from purified particles showed a higher signal at about 7 kb than at 6 kb, suggesting a predominant representation of the full genome molecule in virions compared with that in total DNA (Fig. 6b). At the same short exposure time, two more bands were detected, with slightly fainter signal, at about 4 and 3 kb only in the L34 tDNAex well. These bands were not detected in the encDNA lane (Fig. 6b).

**MBV1-deleted genome variants are present in different mulberry plants**

To investigate the natural occurrence of the deleted shorter MBV1 genome in mulberry, we performed a limited survey...
on mulberry in southern Italy. PCR amplification from total DNA extracts revealed the presence of MBV1 in four out of seven accessions. From each of the virus-positive plants (PV3, B1, MN and Gm), a major fragment (about 2kbp) and a minor one (less than 1kb) were amplified and sequenced (Fig. 7a).

The major band resulted in 2019 nt, aligning with L34-MBV1 between positions 1773 and 3792 and sharing with it 96 %–97 % of sequence identity. When compared to the L34-derived MBV1 sequence, the smaller fragment isolated from all four tested accessions showed a deletion comprised between nt 2563 and 3370, accounting for 807 bp and 92 % overall sequence identity with the correspondent nucleotides of pMBV-7k (Fig. 7a, b). This deletion is smaller than that detected in L34-MBV1 but located in the same genomic context (Fig. 7a, b). Remarkably, the deletion consists of exactly the same genome fragment in viruses retrieved from all mulberry plants analysed from southern Italy, regardless of either species (M. alba or M. nigra) or geographical origin.

The faint band (about 1.2 kb) common to all samples analysed was sequenced and resulted in a fragment of plant transposon.

**DISCUSSION**

The full-length genome sequence of the recently reported MBV1 was obtained either by HTS of sRNAs from the infected mulberry plant or by conventional sequencing of overlapping PCR fragments. Computer-assisted phylogenetic analysis of the MBV1 genome sequence indicates that this virus represents a putatively new species in the Badnavirus genus. The phylogeny of the RT/RNase H domains strictly clusters MBV1 close to the other members of the family Caulimoviridae, in a branch comprising CSSV, CiMV and FBV1. All of the key badnavirus motifs map on the MBV1 sequence, including the tRNA
\[
\text{MET}
\]
which is the primer for the minus-strand synthesis of all Caulimoviridae members (Bouhida et al., 1993; Hohn et al., 1997; Medberry et al., 1990), and the badnavirus-specific DUF1319 (Sether et al., 2012) that is located at the beginning of the single ORF in MBV1. Analysis of the LIGR shows the presence of a stem–loop structure, ending immediately before the main significant coding ORF, in agreement with the hypothesis of ribosomal shunting (Pooggin et al., 1999; Ryabova et al., 2002).

On the other hand, some distinctive features were revealed in the MBV1 genome. A different organization of the gene assembly, compared with the other members of the taxon, is characterized by a single ORF containing a sort of fusion of featuring domains of the ORF1 and 3; and the incomplete structure of the domain DUF1319, having an unknown function, but peculiarly conserved in the ORF1 of all the other badnaviruses. A similar case of a single ORF has previously been described for *Petunia vein clearing virus*...
(PVCV) (Richert-Pöggeler & Shepherd, 1997). In the better-studied RTBV, translation of the different ORFs proceeds similarly from a single, polycistronic pgRNA, which is translated in several proteins either through a frameshift (for smaller ORFs 1 and 2) or protease processing (for the long ORF3) (Fütterer et al., 1997; Ryabova et al., 2006).

A remarkable feature of the MBV1 genome is the coexistence of a less-than-full-length DNA molecule, together with the full-length DNA genome in virus-infected plants. The deleted genome has the intrinsic and stable property of bearing a 1053 bp gap in the region between the MP and the ZF domain with no apparent loss of function. In fact, it

![Diagram of MBV1 genome variants in mulberry plants from southern Italy. (a) Gel image of PCR amplification with specific primer set external to the gap in the shorter virus variant. The red arrow indicates the bands corresponding to the full-length genome (2 kbp) and the blue arrow indicates the bands amplifying from the deleted genome (~1 kbp). WC, water control; PV1, Locorotondo1; PV3, Locorotondo3; MB, Monopoli White; MN, Monopoli Black; B1, Bisceglie1; B3, Bisceglie3; Gm, Mulberry dark Avellino; L34, Lebanon 34. (b) Graphical representation of PCR amplification fragments obtained from the survey. The white box at the top represents ORF1 of MBV1, with relevant genomic features illustrated inside. Illustrated conserved motifs are described at the bottom. Light gray arrows represent primer pairs used for amplification; dark gray lines represent the fragments obtained by PCR amplification, and white boxes illustrate the genome region lacking in the deleted genomes, aligned to the reference. Nucleotide positions are indicated under primers and boundaries of the gap. On the right are indicated the deleted genomes.](image-url)

**Fig. 7.** MBV1 genome variants in mulberry plants from southern Italy. (a) Gel image of PCR amplification with specific primer set external to the gap in the shorter virus variant. The red arrow indicates the bands corresponding to the full-length genome (2 kbp) and the blue arrow indicates the bands amplifying from the deleted genome (~1 kbp). WC, water control; PV1, Locorotondo1; PV3, Locorotondo3; MB, Monopoli White; MN, Monopoli Black; B1, Bisceglie1; B3, Bisceglie3; Gm, Mulberry dark Avellino; L34, Lebanon 34. (b) Graphical representation of PCR amplification fragments obtained from the survey. The white box at the top represents ORF1 of MBV1, with relevant genomic features illustrated inside. Illustrated conserved motifs are described at the bottom. Light gray arrows represent primer pairs used for amplification; dark gray lines represent the fragments obtained by PCR amplification, and white boxes illustrate the genome region lacking in the deleted genomes, aligned to the reference. Nucleotide positions are indicated under primers and boundaries of the gap. On the right are indicated the deleted genomes.
is encapsidated in virus particles and replicates and moves systemically in the host.

Further evidence that the complex replication mechanism of this virus can generate aberrant products is the observation of shorter faint bands in the Southern blot (Fig. 6b), compatible with the existence of replication intermediates produced during the (−)ssDNA synthesis, as previously described for CaMV and RTBV (Bao & Hull, 1994; Thomas et al., 1985; Turner & Covey, 1988).

Bacilliform particles encapsidating the deleted MBV1 genome have a diameter smaller than the wild type and, apparently, a more relaxed capsid structure. While the 1 kb deletion in the MBV1 genome does not hamper virus ability to infect mulberry, it seems to affect the virion morphology. Interestingly, the deletion covers a region of the CP preceding the cysteine-rich ZF domain that mediates the interaction between the CP and the VAP in RTBV (Herzog et al., 2000). Therefore, it is tempting to speculate that the morphological alteration of the genome-deleted MBV1 virion might reflect an impaired (but not abolished) interaction of the CP domain with the VAP-homologue region, which is typically anchored in the virion capsid structure (Plisson et al., 2005).

Deletion events have been described among Caulimoviridae members: a CaMV natural mutant (isolate CM4-184) lacking 421 bp (Howell et al., 1981) and an RTBV Indian subcontinent isolate hosting a deletion of 64 bp (Fan et al., 1996; Sharma et al., 2011). Other shorter or defective genome sequences have been found in the Caulimoviridae family, originating from splicing events in reverse transcripts. However, experimental tests demonstrated that they are neither encapsidated nor able to replicate independently (Füttner et al., 1994; Hirochika et al., 1985; Kiss-László et al., 1995; Scholthof et al., 1991; Vaden & Melcher, 1990). Interestingly, a deletion in the same region as in MBV1, albeit shorter, has recently been found in Grapevine roditis leaf discoloration virus (GRLDV). Two variants of this new badnavirus have recently been described: GRLDV-w4, isolated in Greece (Maliogka et al., 2015), and GRLDV-BN in Italy (Chiumenti et al., 2016b), the first of which presents a discontinuous deletion in the same region missing in the deleted MBV1.

All members of Caulimoviridae replicate via an intermediate RNA, which is retrotranscribed by the viral RT (Bao & Hull, 1994; Hohn et al., 1997). The rearrangement of the genomic DNA and the presence of the deleted genome could be caused by the interaction of two main factors, such as switching and secondary structures of the template. It is known for retroviruses that viral RT is not proofreading and is responsible for the high retrovirus recombination. One of the peculiar recombination mechanisms is indeed the template switching in cis, which means that the RT jumps from one point to another of the template with a significantly higher probability of occurrence of intramolecular swap rather than an intermolecular one (Pathak & Hu, 1997; Temin, 1993). The presence of steady secondary structures in the region of the putative gap could further favour RT switching (Nagy & Simon, 1997; Pathak & Nagy, 2009). Indeed, in silico folding of the MBV1 genomic gap sequence predicted a very stable RNA secondary structure (ΔG=−384.68). Moreover, the invariance of the gap boundaries (hotspots), we determined using sequence analysis of numerous extracts of the L34 MBV1 isolate, further supports the hypothesis of a structural and stably conserved constraint facilitating the synthesis of the deleted virus genome.

A survey in southern Italy identified a different kind of deleted genome in several geographically distant plants, all of them sharing common size and borders of the deletion, coexisting with the parental one. The evidence of a similar evolution or preservation of this feature in different and distant regions of the Mediterranean basin, Italy and Lebanon, suggests that the coexistence of the two (or more) genomic forms might be important for MBV1 fitness, adaptation, or function.

Plant pararetroviruses, including several members of the genus Badnavirus, have the ability to integrate into the host genome (Geering et al., 2000; Stagginus & Richert-Pöggeler, 2006). The latest example is represented by FBV1, a virus phylogenetically related to MBV1, which was shown to occur in both episomal and integrated form in Ficus carica (Laney et al., 2012). In this study, we extracted and sequenced the MBV1 genome from partially purified virus particles, thus confirming the episomal form of the virus. Surveys done on several mulberry plants from different Mediterranean origins (Fig. 7; Elbeaino et al., 2013) revealed that while asymptomatic plants can actually be virus infected, some among the tested accessions clearly show the absence of viral DNA hybridization signal on both Southern blot and PCR. Taken together, these results gather enough clues to assess that MBV1 is not likely to be integrated into the mulberry genome.

Uniquely among the other members of the Caulimoviridae family, MVB1 is characterized by an sRNA profile almost completely lacking the 24-mer size class. While the first sRNA profile of a Badnavirus is reported here, previous studies on sRNAs produced by two DNA viruses, CaMV (genus Caulimovirus) and Cabbage leaf curl virus (CalCuV; genus Begomovirus), revealed that nuclear replicating viruses are characterized by the accumulation of the 24 nt species (Blevins et al., 2006, 2011) generated by DCL3 and possibly involved in de novo methylation of the viral genomic DNA and in regulation of plant antiviral defence (Pooggin, 2013; Raja et al., 2010). The evidence is that sRNAs isolated from CaMV-infected A. thaliana mutants defective for DLC2/DCL3/DCL4 show a profile lacking virus-derived 24-mers (Blevins et al., 2011), which prompts the speculation that MBV1 could escape the action of DCL3 to avoid de novo DNA methylation.

Based on the hypothesis that plant pararetroviruses can escape methylation pathways by delivering to the nucleus unmethylated copies of circular dsDNA produced in the
cytoplasm from the pgRNA (Pooggin, 2013), it is possible to hypothesize that the MBV1-deleted genome is less abundant in infected plants because it acts as a silencing decoy, thus facilitating the replication of the parental genome. Alternatively, most of the plant defence activity against MBV1 would take place in the cytoplasm as post-transcriptional repression during DNA genome replication. However, further experimental work is needed to demonstrate any of these hypotheses.

Sorting of sRNAs into a certain AGO protein is a process dependent on their length and on the nucleotide at the 5’ terminus. While the MBV1 genome is prevalently composed of A and T nucleotides (33.7 % and 21.8 %, respectively), the frequency of these specific residues at virus-derived sRNA 5’ terminal positions is not directly dependent on their abundance in the virus genomic DNA. This supports the implication of additional factors, including selective AGO recruiting (Kim, 2008; Mi et al., 2008).

Undoubtedly, MBV1 has proved to be a new and interesting virus species deserving deeper investigation of its functional organization and the plant defence mechanisms it can trigger. The interesting finding, that deletion of large portions of the MBV1 genome may produce a viable virus, queries the minimal size of an autonomously replicating genome and opens new evolutionary scenarios of such divergent behaviour in the badnavirus genus.

**METHODS**

**Plant material.** Source plant of MBV1 was the accession Lebanon 34 (L34) of *M. alba* L., sampled in Lebanon in 2011 (Elbeaino et al., 2013), which shows symptoms of leaf mottling and vein yellowing. A small survey for the molecular detection of MBV1 was completed in Apulia and Campania (Southern Italy), selecting a few apparently symptomless field-grown *Morus* trees.

**Preparation of sRNA libraries and NGS sequence analysis.** Total RNA was extracted in late spring from 1 g of symptomatic leaf tissue using the TRizol reagent (Life Technologies), prior to low-molecular-weight fraction recovery by polyethylene glycol precipitation (Hamilton & Baulcombe, 1999). sRNAs of 19–26 nt in length were separated by 15 % denaturing polyacrylamide gel and recovered by excision (Lu et al., 2007). Libraries from recovered sRNAs were prepared for sequencing on HiScanTM SQ (Illumina), using TruSeqTM Small RNA Sample Preparation kit (Illumina).

The cDNA library was submitted to a 50-base single-read run. Short sequences obtained were pre-processed with fastQ toolkit, de novo assembled with Velvet (Zerbino & Birney, 2008) and searched for homology sequences by BLASTN/BLASTX (Altschul et al., 1990). Guided assembly of the viral sequences was performed with SOAP (Li et al., 2008), with two default mismatches, using the reference sequence of *Fig badnavirus*1 (FBV1; GenBank accession no. NC017830), *Cacao swollen shoot virus* (CSSV; NC001574) and *Citrus yellow mosaic virus* (CYMV; NC003382) and MBV1-assembled sequence, respectively. The choice of these viruses was suggested by previous phylogenetic analysis (Elbeaino et al., 2013). Alignment, using FBV1 as a reference sequence was undertaken to identify the most conserved regions, including any possible nucleotide polymorphism beyond the first-obtained contigs. To obtain the MBV1 full-length sequence, some primers were designed in conserved regions from the contigs aligned to FBV1 genome.

**Sequencing strategies.** Total nucleic acids from 300 mg of L34 leaves were extracted according to Murray & Thompson (1980). Circular DNA enrichment was performed on Eurogold columns (Euroclone). This template was used for PCR and inverse-PCR experiments with 14 different primer sets. TaKaRa SpeedSTAR™ (Takara Bio) was applied for PCR with a two-step amplification protocol. PCR amplicons of the expected sizes were purified and directly sequenced (Macrogen Europe).

**Gene structure and homology search on MBV1 genome.** Conventional sequences from PCR products were assembled using the BioEdit 7.0.9 program CAP contig assembly tool (Hall, 1999). ORF identification was obtained with ORF Finder (NCBI) and a study of conserved domains was done through Conserved Domain Database with CD-Search (NCBI; Marchler-Bauer et al., 2008). Alignments for nucleotide and amino acid homology were done with the conserved program (Larkin et al., 2007). Similarity to known proteins from the protein information resources was determined using BLASTx and BLASTn (Altshul et al., 1997) algorithms. Phylogenetic relationships among badnaviruses were evaluated with the neighbour-joining method, using the MEGAB program (Tamura et al., 2013) with 1000 replicates of bootstrap.

**In silico validation of conventional sequencing results and sRNA characterization.** In silico validation of conventional sequencing results and sRNA characterization was carried out running a new SOAP alignment, using the MBV1 assembled sequence as a reference. Secondary structures of putative pregenomic RNA, delimited according to Pooggin et al. (1999), were scanned using the mFold web server (Zucker, 2003). In silico coiled-coil identification was performed with the Coils tool on the ExPASy online platform, based on the method described by Lupas et al. (1991, 1996).

**Infectious clone synthesis and mechanical inoculation of mulberry seedlings.** Total DNA (tDNAex) was extracted from L34 leaves and the circular DNA fraction was enriched as described above. An increased amount of specific template was obtained by RCA (Dean et al., 2001) with the Illustra TempliPhi Amplification kit (GE Healthcare Life Science), adding a mix of 12 specific MBV1 primers (final concentration 0.83 µM each).

Two adjacent primers, SpeI-F and SpeI-R, were designed on the *Spel* unique restriction site of the full-length sequence, with the aim of amplifying the MBV1 complete genome. Therefore, RCA dilutions (up to 1:200) were used as template for SpeI-F/R amplification with Phusion High-Fidelity DNA Polymerase (Thermo Scientific). Two PCR products obtained, of about 7 kbp and 6 kbp, were cloned with the StataClone PCR Cloning Kit (Agilent Technologies). Several obtained clones were screened by PCR amplification with different specific primer sets (in the RNase H and DUF1319 regions); two clones of different size, ~6 and 7 kb (named pMBV-6k and pMBV-7k) were finally identified and completely sequenced (Macrogen Europe).

Since the described amplification and cloning apparently also selected a less-than-full-length, deleted genome (see Results), we needed to confirm the existence of both molecular forms in planta. Therefore, a set of primers was designed on two unique restriction sites (*BstI* and *BpiI*; see Fig. 1a) flanking the identified putative gap. These primers (Gap-for/Gap-rev) were used on the L34 tDNAex in order to amplify, eventually, two differently sized bands on the templates.

DNA preparations of both clones (pMBV-6k and pMBV-7k), purified on QIAGEN Midi Prep columns (Qiagen, The Netherlands) at a concentration of 20 µg µl⁻¹, were digested with *Sphel* and used for inoculations. Eighteen mulberry seedlings (6 *M. alba* and 12 *M. nigra*) were selected. Two mature pairs of true leaves were inoculated by mechanical abrasion with plasmid pMBV-6k and pMBV-7k, while 5 additional seedlings were rubbed with buffer only. Next, 10 µg of digested plasmid was spread on QIAGEN Midi Prep columns (Qiagen, The Netherlands) at a concentration of 20 µg µl⁻¹, were digested with *Sphel* and used for inoculations. Eighteen mulberry seedlings (6 *M. alba* and 12 *M. nigra*) were selected. Two mature pairs of true leaves were inoculated by mechanical abrasion with plasmid pMBV-6k and pMBV-7k, while 5 additional seedlings were rubbed with buffer only. Next, 10 µg of digested plasmid was spread onto the surface of each of the two inoculated leaves per seedling. Plants were indefinitely kept under greenhouse conditions (25 °C; 16 h light period). Several seedling batches from different mother plants, infected or not by MBV1, invariably tested negative for virus detection by PCR.
Inoculated plants were periodically inspected for symptom expression. Symptoms eventually occurring on plants inoculated with both constructs were compared to those visible on the L34 original accession and to mock-inoculated plants. At 12 weeks post-inoculation, DNA was extracted from 100 mg of systemic leaf tissues (Murray & Thompson, 1980). DNA was quantified and diluted to 100 ng µl⁻¹ for PCR amplifications. Primer pairs in RNase H and DUF1319 regions were used for detection.

For EM analysis, each leaf sample was excised from plants and ground in phosphate buffer, then a droplet was applied to a carbon-coated copper/rhodium grid (Milne, 1993). The specimen, rinsed and negatively stained, was examined by a Philips Morgagni 282D electron microscope operating at 60 kV. To determine the length and diameter of the observed virion morphologies, metadata were extrapolated from scanned frames using the software Image 1 : 3 : 38 (Copyright 1999–2004, Gisle Aas) and statistically evaluated by the Kalexigraph software 3.0 (Synergy Software).

Finally, DNA from a plantlet (p5) inoculated with the clone pMBV-6k, PCR-positive to MBV1 infection, was used to re-sequence the virus genome. Different combinations of primers spanned along the virus genome, were gel eluted, cloned and sequenced.

**Virus purification and analysis of encapsidated DNA.** Twenty grams of fully expanded, symptomatic leaves from the L34 plant were collected and ground in liquid nitrogen. The powder was stirred overnight at 4 °C in the extraction buffer (0.2 M Tris-HCl, pH7; 0.02 M EDTA; 1.5 M Urea, 2 % Triton X-100) at the ratio of 10 ml g⁻¹ leaf tissue and further processed according to the method used by Hull et al. (1976).

After a final DNase I digestion on purified virion suspension, followed by a proteinase-K treatment and phenol/chloroform extraction, encapsidated DNA was checked by PCR using primers in the RNase H region and Gap-for/Gap-rev. An inverse PCR was performed on this purified virus DNA, with primers PCR80-for/PCR80-rev. Obtained amplicons were gel eluted, cloned and sequenced.

Tenfold serial dilutions were prepared with (i) the two constructed clones, pMBV-6k and pMBV-7k; (ii) the encapsidated DNA from purification (encDNA); and (iii) the tDNA from L34. A concentration of about 700 ng µl⁻¹ obtained by optical density reading (Nanodrop, Thermo Scientific) was the starting point for both plasmids, and the dilution endpoints were 10⁻⁹.

Southern blot hybridization was performed to verify the presence of either the full-size genome or the putative deleted DNA in the L34 original accession, and to determine whether the virus DNA was integrated into the mulberry genome or replicated as an episomal form. DNA from mulberry plants was extracted as previously described (Murray & Thompson, 1980) and digested with SpeI and RNase A. Equal amounts of DNA extract (1–2 µg each) were electrophoresed in a 0.7 % agarose gel and transferred onto a Hybond N+ membrane (GE Healthcare). EncDNA (200 ng µl⁻¹) from partially purified virions was treated in the same way. Four MBV1 amplicons, spanning along the whole genome, were labelled on both strands during PCR by digoxigenin-11-dUTP (Roche Applied Science). Hybridization was performed at 46 °C overnight and then the membrane was subjected to chemiluminescent detection.

**ACKNOWLEDGEMENTS**

We thank Professor Benham E. L. Lockhart (Department of Plant Pathology, University of Minnesota) for critical reading of the manuscript. The NGS activity was hosted by the Illumina platform at ’Rete di laboratori pubblici di ricerca’ SELGE. We also thank Arborea sementi (Italy) for providing seeds of *M. alba* and *M. nigra*.

**REFERENCES**


