Acta virologica 67: 13 – 23, 2023 doi: 10.4149/av_2023_102

High-throughput RNA sequencing analysis of *Mallotus japonicus* revealed novel polerovirus and amalgavirus

Dongjin Choi[1], Megha Rai[2,3], Amit Rai[3,4], Chaerim Shin[1], Mami Yamazaki[2,3], Yoonsoo Hahn[1*]

[1]Department of Life Science, Chung-Ang University, Seoul 06974, South Korea; [2]Graduate School of Pharmaceutical Sciences, Chiba University, Chiba 260-8675, Japan; [3]Plant Molecular Science Center, Chiba University, Chiba 260-8675, Japan; [4]RIKEN Center for Sustainable Resource Science, Yokohama, Kanagawa 230-0045, Japan

Received September 09, 2022; revised November 09, 2022; accepted November 14, 2022

Summary. – High-throughput RNA sequencing (RNA-seq) analysis of samples from *Mallotus japonicus*, a traditional medicinal plant, yielded two novel RNA viruses tentatively named Mallotus japonicus virus A (MjVA) and Mallotus japonicus virus B (MjVB). The MjVA and MjVB genomes encode proteins showing amino acid sequence similarities to those of poleroviruses (the genus *Polerovirus*, the family *Solemoviridae*) and amalgaviruses (the genus *Amalgavirus*, the family *Amalgaviridae*), respectively. The MjVA genome contains seven highly overlapping open reading frames, which are translated to seven proteins through various translational mechanisms, including –1 programmed ribosomal frameshifting (PRF) at the slippery motif GGGAAAC, non-AUG translational initiation, and stop codon readthrough. The MjVB genome encodes two proteins; one of which is translated by +1 PRF mechanism at the slippery motif UUUCGN. The abundance analysis of virus-derived RNA fragments revealed that MjVA is highly concentrated in plant parts with well-developed phloem tissues as previously demonstrated in other poleroviruses, which are transmitted by phloem feeders, such as aphids. MjVB, an amalgavirus generally transmitted by seeds, is distributed in all samples at low concentrations. Thus, this study demonstrates the effectiveness and usefulness of RNA-seq analysis of plant samples for the identification of novel RNA viruses and analysis of their tissue distribution.

Keywords: Polerovirus; Amalgavirus; Mallotus japonicus; RNA virus; viral genome; programmed ribosomal frameshifting

Introduction

The high-throughput sequencing of genetic materials acquired from diverse host organisms and environments has greatly broadened our knowledge on the diversity of viruses (Shi *et al*., 2016). The comprehensive analysis of RNA sequencing (RNA-seq) data obtained primarily for gene expression studies has also facilitated the detection and characterization of RNA viruses (Kim *et al*., 2014; Edgar *et al*., 2022). RNA samples isolated from cellular organisms, especially plants, may contain RNA molecules derived from RNA viruses. Many novel RNA virus genome sequences have been identified by analyzing RNA-seq data from plant samples (Choi *et al*., 2021, 2022; Shin *et al*., 2021, 2022).

Mallotus japonicus is a valuable traditional medicinal plant widely distributed in East Asia; its tissues, such as leaves, root, and bark, have been used to treat several diseases, including stomach disorders, irritable bowel syndrome, rheumatism, diabetes, and neuralgia (Arisawa, 1994; Wu *et al*., 2021). High-throughput RNA-seq

*Corresponding author. E-mail: hahny@cau.ac.kr; phone: +82-2-820-5812.

Abbreviations: FPKM = fragments per kilobase per million; MjVA = Mallotus japonicus virus A; MjVB = Mallotus japonicus virus B; NLS = nuclear localization signal; ORF = open reading frame; PRF = programmed ribosomal frameshifting; RdRp = RNA-dependent RNA polymerase; SaYV = Sauropus yellowing virus; SRA = Sequence Read Archive

and metabolite profiling of seven tissues of *M. japonicus* have revealed that metabolite accumulations are strongly correlated with gene expression among analyzed tissues (Rai *et al.*, 2021).

Poleroviruses are plant pathogenic viruses that cause quality and yield losses of economically important crop plants (Distefano *et al.*, 2010; Delfosse *et al.*, 2021). The genus *Polerovirus* belongs to the family *Solemoviridae*, together with three other genera, namely, *Enamovirus*, *Polemovirus*, and *Sobemovirus* (Sõmera *et al.*, 2021). Poleroviruses have a positive-sense single-stranded RNA genome of 5–6 kb in length and contain seven conserved overlapping open reading frames (ORFs), namely, ORF0, ORF1, ORF2, ORF3a, ORF3, ORF4, and ORF5 (LaTourrette *et al.*, 2021; Igori *et al.*, 2022). These ORFs produce proteins called P0 (ORF0), P1 (ORF1), P1–P2 (fusion of ORF1 and ORF2), P3a (ORF3a), P3 (ORF3), P3–P5 (fusion of ORF3 and ORF5), and P4 (ORF4). The P1–P2 fusion protein, which contains an RNA-dependent RNA polymerase (RdRp) domain, is generated by −1 programmed ribosomal frameshifting (PRF), which occurs at the conserved slippery heptanucleotide sequence GGGAAAC within ORF1 (Atkins *et al.*, 2016; Delfosse *et al.*, 2021). A pseudoknot structure after this consensus is required for −1 PRF (Csaszar *et al.*, 2001; Atkins *et al.*, 2016; Delfosse *et al.*, 2021). P3a protein translation is initiated at a non-AUG start codon (Smirnova *et al.*, 2015). The P3–P5 fusion protein is produced via the readthrough translation of the UAG stop codon of ORF3 (Knierim *et al.*, 2015; LaTourrette *et al.*, 2021). Several aphid species serve as transmission vectors of most poleroviruses, while a whitefly species acts as a vector of some poleroviruses (Ghosh *et al.*, 2019; LaTourrette *et al.*, 2021).

Amalgaviruses are members of the family *Amalgaviridae*, which is composed of two approved genera (*Amalgavirus* and *Zybavirus*) and a proposed genus "Anlovirus" (Krupovic *et al.*, 2015; Depierreux *et al.*, 2016; Pyle *et al.*, 2017). They are vertically transmitted through seeds from one generation to the next with or without causing visible symptoms (Martin *et al.*, 2011). They have a double-stranded RNA genome with a length of approximately 3.5 kb and encode two ORFs, namely, ORF1 and ORF2 (Park and Hahn, 2017; Park *et al.*, 2018). Although the ORF1 product or ORF1p has no established function, it likely participates as a nucleocapsid or replication factory-like protein (Isogai *et al.*, 2011; Krupovic *et al.*, 2015). ORF2 encodes RdRp and is translated as the ORF1+2p fusion protein; this process is mediated by +1 PRF mechanism occurring at the slippery sequence UUUCGN, where N is any nucleotide (Nibert *et al.*, 2016; Goh *et al.*, 2018; Lee *et al.*, 2019).

In the present study, a novel polerovirus and a novel amalgavirus were identified via the high-throughput RNA-seq of tissue samples from *M. japonicus*.

Materials and Methods

RNA-seq data. Seven samples, including young leaves, mature leaves, young stems, mature stems, bark, central cylinder, and inflorescence, of a 12-year-old *M. japonicus* plant were collected for high-throughput RNA-seq analysis (Rai *et al.*, 2021). RNA-seq data are available in the Sequence Read Archive (SRA) of the National Center for Biotechnology Information (NCBI) under the Acc. Nos. SRR15027072–SRR15027078. Raw *M. japonicus* RNA-seq data were filtered to obtain high-quality reads by using sickle (version 1.33; https://github.com/najoshi/sickle) with the parameter "-q 30 -l 55." High-quality reads were assembled into contigs by using the SPAdes Genome Assembler (version 3.15.4; http://cab.spbu.ru/software/spades) with the parameter "--rnaviral," which is optimized for the generation of RNA viral genome contigs (Bushmanova *et al.*, 2019).

Identification of viral genome contigs. Putative viral genomes in the assembled *M. japonicus* RNA-seq contigs were initially identified by comparing them with known viral RdRp sequences via the BLASTX mode of DIAMOND (Buchfink *et al.*, 2021). Known viral RdRp domain sequences were obtained from the Pfam database (release 35.0; https://pfam.xfam.org; Pfam Acc. Nos.: PF00602, PF00603, PF00604, PF00680, PF00946, PF00972, PF00978, PF00998, PF02123, PF03431, PF04196, PF04197, PF05273, PF05788, PF05919, PF06317, PF06478, PF07925, PF08467, and PF12426). Putative viral genome contigs were compared with all known viral proteins by using the NCBI BLAST server (https://blast.ncbi.nlm.nih.gov/Blast.cgi) to select novel virus genome sequences.

Annotation and analysis of viral genomes. The ORFs of a viral genome contig were initially predicted using the ORF finder web server (https://www.ncbi.nlm.nih.gov/orffinder). Special features, such as a PRF site, non-AUG start codon, and stop codon readthrough, were determined by comparing the nucleotide sequences with those of closely related viruses. A pseudoknot structure was predicted using the IPknot++ web server (http://rtips.dna.bio.keio.ac.jp/ipknot++) (Sato *et al.*, 2011). Signal peptides and transmembrane domains were predicted using SignalP (version 6.0; https://services.healthtech.dtu.dk/service.php?SignalP) and DeepTMHMM (version 1.0.12; https://dtu.biolib.com/DeepTMHMM), respectively (Hallgren *et al.*, 2022; Teufel *et al.*, 2022). A nuclear localization signal (NLS) was predicted via NLStradamus (version r.9; http://www.moseslab.csb.utoronto.ca/NLStradamus) (Nguyen Ba *et al.*, 2009). The BOXSHADE program (version 3.31; https://launchpad.net/ubuntu/focal/+package/boxshade) was used to visualize multiple sequence alignments of PRF regions. The WebLogo web server (version 3; http://weblogo.threeplusone.com) was used to create sequence logo representation (Schneider and Stephens, 1990; Crooks *et al.*, 2004).

Phylogenetic analysis. The genome and protein sequences of poleroviruses and amalgaviruses were collected by sequence-similarity and keyword searches of the NCBI sequence data-

bases. The multiple alignments of viral protein sequences were generated by using MAFFT (version 7.490; https://mafft.cbrc.jp/alignment/software) with the parameter "--auto" (Nakamura *et al.*, 2018). Multiple sequence alignments were filtered using trimAl (version 1.4.rev22; http://trimal.cgenomics.org), with the parameter "-automated1," to select well-aligned informative positions optimized to reconstruct a maximum likelihood phylogenetic tree (Capella-Gutiérrez *et al.*, 2009). Maximum likelihood trees were inferred using IQ-TREE (version 2.1.3; http://www.iqtree.org) (Minh *et al.*, 2020). Bootstrap support values were calculated from 1000 replicates by using the UFBoot2 method implemented in the IQ-TREE program (parameter "-B 1000"). Phylogenetic trees were visualized using MEGA (version 11.0.11; https://www.megasoftware.net) (Tamura *et al.*, 2021).

Viral read abundance analysis. The abundances of virus-derived reads in *M. japonicus* RNA-seq data were calculated by mapping high-quality reads to viral genome sequences by using the BWA-MEM algorithm of the BWA Aligner (version 0.7.17-r1194-dirty; https://github.com/lh3/bwa). SAMtools (version 1.14; https://github.com/samtools/samtools) was used to extract RNA-seq reads that mapped to viral genomes (Danecek *et al.*, 2021). For normalization, fragments per kilobase per million (FPKM) values were calculated according to the following formula: $10^9 \times C / (N \times L)$, where C is the number of fragments mapped onto the virus genome, N is the total number of high-quality fragments, and L is the length of a virus genome.

Results and Discussion

Identification of novel RNA virus genomes

RNA-seq data, a total of 16 gigabases, were generated from seven samples (young leaves, mature leaves, young stems, mature stems, bark, central cylinder, and inflorescence) of a 12-year-old *M. japonicus* plant to study gene-metabolite networks involving tissue-specific accumulation of therapeutically important substances (Rai *et al.*, 2021). When *M. japonicus* RNA-seq assembled contigs were compared with representative known RNA viral RdRp sequences, two contigs, namely, 6217 and 3362 nucleotides (nt) in length, showed strong sequence similarities to known viral RdRp sequences. Subsequent sequence similarity searches of all known proteins in the NCBI confirmed that these two contigs were nearly complete viral genomes of novel RNA viruses. They were tentatively named Mallotus japonicus virus A (MjVA, a 6217 nt long contig) and Mallotus japonicus virus B (MjVB, a 3362 nt long contig). The genome sequences of MjVA and MjVB were deposited in the NCBI nucleotide database under Acc. Nos. OP122168 and OP122169, respectively.

MjVA is a novel polerovirus

The MjVA genome was 6217 nt long and predicted to encode an RdRp and other proteins most closely related with those of poleroviruses, including carrot red leaf virus, cotton leafroll dwarf virus, Plantago asiatica virus A, and Sauropus yellowing virus (SaYV) (Huang *et al.*, 2005; Distefano *et al.*, 2010; Knierim *et al.*, 2015; Igori *et al.*, 2022). The MjVA genome was predicted to have seven ORFs, which are translated to seven proteins, including P0 (ORF0), P1 (ORF1), P1–P2 (fusion of ORF1 and ORF2), P3a (ORF3a), P3 (ORF3), P3–P5 (fusion of ORF3 and ORF5), and P4 (ORF4; Fig. 1a). No additional ORFs, which have been found in some poleroviruses, were predicted in the MjVA genome (LaTourrette *et al.*, 2021).

The MjVA ORF0 is located at the 116–910 genome position, which encodes the 264-amino acid (aa) protein P0. As in other poleroviruses, most of the MjVA ORF0 sequence overlaps with the ORF1 at the 282–2483 position. The polerovirus P0 protein is a suppressor of RNA silencing and interferes with the RNA silencing defense mechanism of the plant hosts (Delfosse *et al.*, 2021).

The MjVA ORF1 (282–2483 genome position) encodes 733 aa P1 protein, which was predicted to have a signal peptide at the N-terminus, at the 1–18 aa position, and a transmembrane domain at the 147–148 aa position. These hydrophobic segments may mediate the attachment of the P1 protein to intracellular membranes as observed in other polerovirus P1 proteins (Delfosse *et al.*, 2021).

The MjVA ORF2 (1961–3748 genome position) contains an RdRp domain, which is required for the viral genome replication. The polerovirus ORF2 is translated as a P1–P2 fusion protein by –1 PRF mechanism. At a low rate, during ORF1 translation, ribosomes may move backward by one nucleotide at the conserved slippery heptanucleotide sequence GGGAAAC, which is followed by a pseudoknot structure (Csaszar *et al.*, 2001; Atkins *et al.*, 2016; Delfosse *et al.*, 2021). The putative slippery sequence GGGAAAC was found in the MjVA genome at the 1955–1961 genome position, which was followed by a possible pseudoknot structure (Fig. 1b). The C residue located at the 1961 position would be the first base of ORF2. Therefore, the MjVA P1–P2 fusion protein, 1155 aa in length, is produced by translating a part of ORF1 (282–1961 position) and the entire ORF2 (1961–3748 position). Nucleotide residues at the 1968–1972 position (5'-GGCCG-3') and those at the 1979–1983 position (3'-CCGGC-5') might form the first stem of the pseudoknot structure. The second stem was predicted to be formed between residues at the 1972–1978 position (5'-GCUGG-3') and those at the 1997–1993 position (3'-CGACC-5'). Another conserved sequence motif AAACAA, which is shared among poleroviruses, was

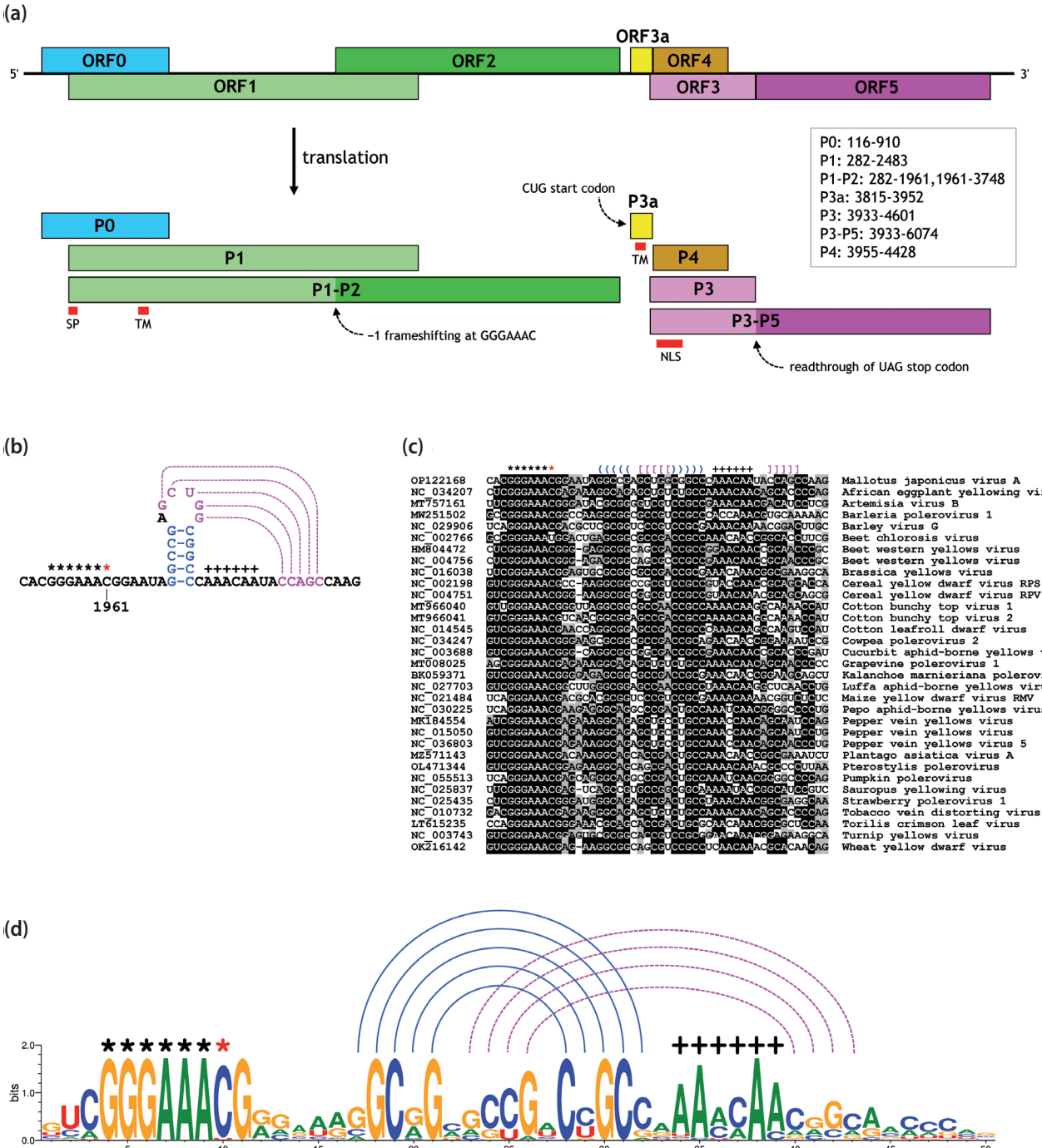

Fig. 1

Genomic organization of Mallotus japonicus virus A (MjVA)

(a) Schematic representation of the MjVA genome is shown. ORFs and protein products are depicted as differentially colored boxes. Thin red boxes indicate a predicted signal peptide (SP) for P1 and P1–P2 proteins; transmembrane domains (TM) of P1, P1–P2, and P3a proteins; and a nuclear localization signal (NLS) of P3 and P3–P5 proteins. The genomic positions of protein-coding regions are presented in a box at the right. **(b)** Predicted pseudoknot structure involving -1 programmed ribosomal frameshifting (PRF) is shown. Residues forming two stem structures are indicated by blue and magenta, respectively, and are connected by lines in their corresponding colors. **(c)** Multiple alignment of sequences surrounding the -1 PRF sites of MjVA and representative poleroviruses is shown. Residues forming stems are indicated by parenthesis and square brackets. **(d)** Sequence logo representation of the multiple alignment of the -1 PRF sites is presented. In **(b)**, **(c)**, and **(d)**, the slippery sequence GGGAAAC and the conserved sequence AAACAA are marked with asterisks and plus signs, respectively. The red asterisk indicates the first nucleotide of ORF2.

found at the 1985–1990 position within the second loop of the pseudoknot structure (Delfosse *et al.*, 2021).

Multiple alignment comparison of nucleotide sequences around the –1 PRF site revealed a strong sequence conservation of slippery heptanucleotide sequences and pseudoknot-forming regions among poleroviruses (Fig. 1c). The sequence logo representation of the alignment provided a clear view of the detailed sequence conservations (Fig. 1d). In addition to the slippery heptanucleotide sequence GGGAAAC, the G residue next to the motif was extremely conserved, suggesting that the slippery motif is the octanucleotide sequence GGGAAACG. The first stem is formed by highly conserved nucleotide residues of which deduced consensus sequences are 5'-DGCRG-3' and 3'-VCGYC-5', where D represents A, G, or U; R represents A or G; V represents A, C, or G; and Y represents C or U. The second stem also showed a moderate sequence conservation with the most frequent sequences being 5'-GCCG-3' and 3'-CGGC-5'. Another conserved motif is found within the second loop, and the most common sequence is AAACAA. However, the exact locations and base-pairing patterns of pseudoknot structures may vary among poleroviruses. Nonetheless, these conserved sequence motifs can be utilized to infer the general features of polerovirus –1 PRF sites and predict them from novel polerovirus genome sequences.

The MjVA ORF3a is located at the 3815–3952 genome position, encoding a 45 aa P3a protein, with a CUG codon as the predicted start codon. All known polerovirus ORF3a start with a non-AUG codon. When 36 known polerovirus genome sequences with annotated ORF3a were analyzed, the most frequent ORF3a start codon was AUA with 20 cases, followed by ACG, 6; AUU, 5; AUG, 2; CUG, 2; and GUG, 1. The MjVA P3a was predicted to have a transmembrane domain at the 6–26 aa position, which may be required for its plasma membrane localization (Smirnova *et al.*, 2015; Delfosse *et al.*, 2021).

The MjVA ORF3 is located at the 3933–4601 genome position and encodes 222 aa P3 protein. ORF3 starts within ORF3a and two ORFs share 20 nucleotides in different reading frames. The polerovirus P3 protein is the coat protein that encapsidates genomic RNA molecules (Delfosse *et al.*, 2021). An arginine-enriched basic region was found in the MjVA P3 protein at the 16–68 aa position, with two arginine-rich segments at 16–27 (RRRRNRRRRQRR) and 55–64 (RRRRRRNRRR) positions. This segment was predicted to be an NLS, which may be responsible for the nuclear localization of MjVA P3 (Haupt *et al.*, 2005).

The stop codon of polerovirus ORF3 is not 100% efficient, and in some instances, ribosomes continue to translate the succeeding ORF5 by incorporating a glutamine, tyrosine, or histidine residue instead of terminating the process; as a result, the P3–P5 fusion protein is produced (Xu *et al.*, 2018). The P3–P5 fusion protein is a minor coat protein assembled into viral particles (Peter *et al.*, 2008). The MjVA ORF5 is located at the 4602–6074 genome position after ORF3. The readthrough translation of the MjVA ORF3 stop codon results in 713 aa P3–P5 fusion protein translated from the 3933–6074 genome position. All the 36 known poleroviruses and MjVA have UAG as the ORF3 stop codon.

The MjVA ORF4 is entirely embedded within ORF3 in different reading frame and separated from ORF3a by only two nucleotides, spanning at the 3955–4428 genome position. The tightly packed organization of ORF3a, ORF3, and ORF4 is a common feature of known poleroviruses (LaTourrette *et al.*, 2021). The MjVA ORF4 encodes 157 aa P4 protein, which participates in cell-to-cell movement through plasmodesmata and systemic long-distance movement via phloem sieve elements (Delfosse *et al.*, 2021).

The phylogenetic relationship of MjVA and known poleroviruses was investigated using the P1–P2 fusion protein sequences (Fig. 2). A maximum likelihood phylogenetic tree was inferred from P1–P2 fusion protein sequences of MjVA, 72 poleroviruses, and 5 enamoviruses (the genus *Enamovirus*, the family *Solemoviridae*). It revealed that MjVA is a distinct member of the genus *Polerovirus*. The closest known member was SaYV, which formed a subclade with MjVA with a bootstrap value of 93% (Knierim *et al.*, 2015). The MjVA and SaYV P1–P2 proteins showed 46% pairwise aa sequence identity over 989 aligned residues.

MjVB is a novel amalgavirus

The second virus identified in *M. japonicus* RNA-seq data was MjVB with a genome of 3362 nt. Sequence database searches revealed that the MjVB genome contained an RdRp domain similar to those of amalgaviruses, including Cleome droserifolia amalgavirus 1 (CdAV1), Gevuina avellana amalgavirus 1, Medicago sativa amalgavirus 1 (MsAV1), and Vicia cryptic virus M (VCV-M) (Nibert *et al.*, 2016; Zhang *et al.*, 2020). The MjVB genome sequence has two ORFs, namely, ORF1 and ORF2, whose organization is universally observed in all known amalgaviruses (Fig. 3a) (Park and Hahn, 2017; Park *et al.*, 2018). These two ORFs are translated to two proteins known as ORF1p and ORF1+2p fusion proteins.

The MjVB ORF1 is located at the 78–1289 genome position and encodes 403 aa ORF1p protein. The function of amalgavirus ORF1p is yet to be established although it may serve as a nucleocapsid or replication factory-like protein (Isogai *et al.*, 2011; Krupovic *et al.*, 2015).

The amalgavirus ORF2 has an RdRp domain and is translated as the ORF1+2p fusion protein by +1 PRF

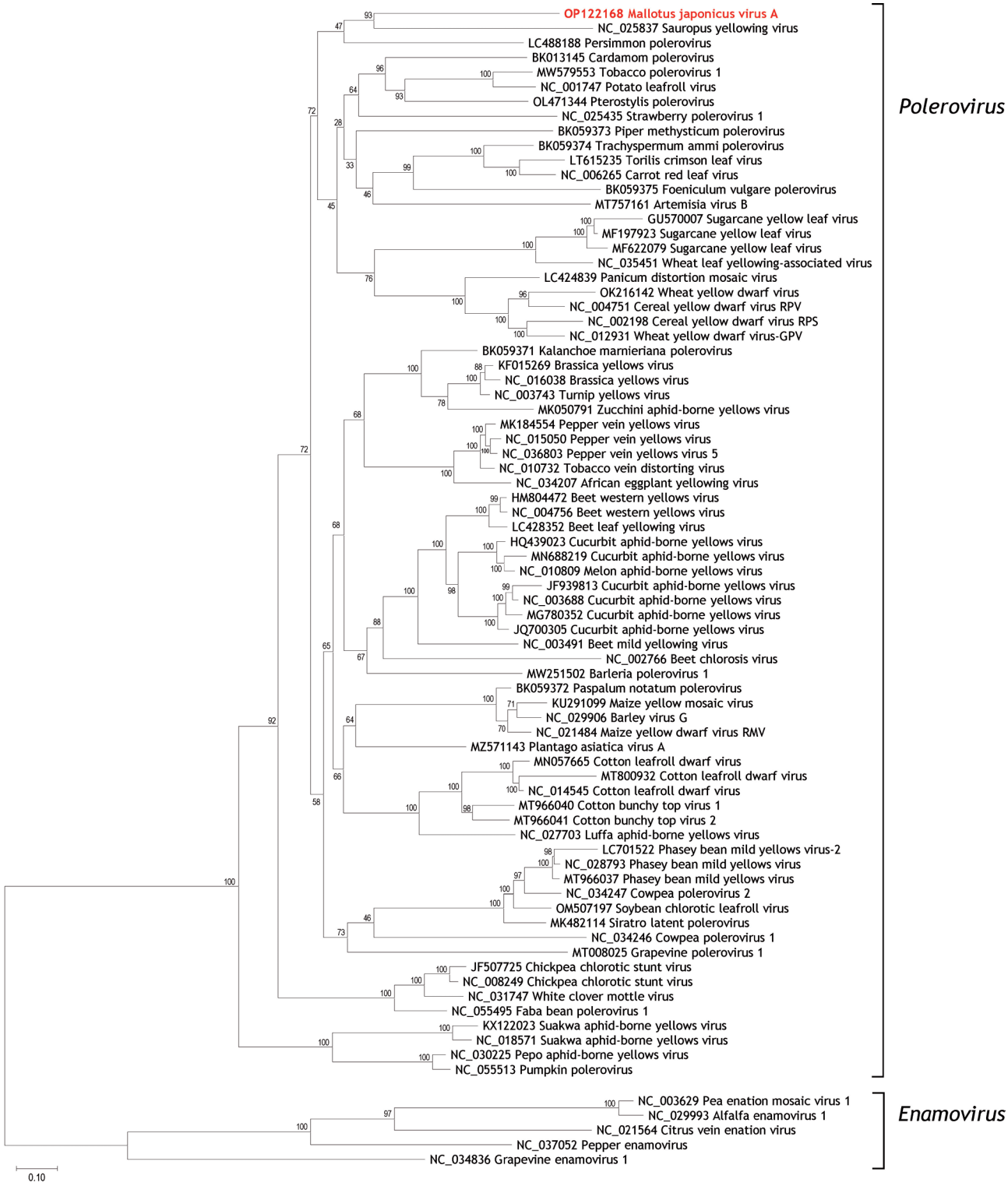

Fig. 2

Phylogenetic analysis of Mallotus japonicus virus A (MjVA)

The P1–P2 fusion protein sequences of MjVA, 72 poleroviruses, and 5 enamoviruses were used to prepare a maximum likelihood phylogenetic tree. Note that MjVA and Sauropus yellowing virus form a clade with a bootstrap value of 93%. Bootstrap values, calculated from 1,000 replicates, were shown at the nodes. The *Enamovirus* clade was used as an outgroup.

mechanism (Nibert *et al.*, 2016; Goh *et al.*, 2018; Lee *et al.*, 2019). The probable slippery sequence UUUCGG was found in the MjVB genome at the 972–979 genome position. The C residue located at 975 is skipped during ORF1 translation, which results in ORF2 translation. Therefore, the MjVB ORF1+2p fusion protein, 1070 aa in length, is produced by the fusion of a part of ORF1 (78–974 position) and the whole ORF2 (976–3291 position).

Multiple alignment comparison of the amalgavirus +1 PRF site sequences confirmed the strong conservation of slippery hexanucleotide sequences (Fig. 3b). The sequence logo of the alignment showed that the slippery consensus UUUCGN was efficiently conserved (Fig. 3c). The most common nucleotide for the N position was the U residue. The C residue located at three nucleotides downstream of the slippery hexanucleotide site also showed an increased frequency, suggesting its possible role during +1 PRF.

The phylogenetic position of MjVB was investigated using the ORF1+2p protein sequences of MjVB and known amalgaviruses. The MjVB ORF1+2p protein showed similar sequence identities with those from many known amalgaviruses with approximately 45% aa identity over about 1,000 aligned residues. This observation implied that MjVB was a novel amalgavirus with no strong affinity to any known amalgaviruses. A maximum likelihood phylogenetic tree inferred from multiply aligned ORF1+2p protein sequences confirmed this speculation (Fig. 4). The MjVB formed a clade with the subclade consisting of VCV-M, MsAV1, and CdAV1. However, its bootstrap value was only 54%, indicating that the relatedness of MjVB and these three amalgaviruses was poorly supported. Therefore, MjVB is a distinct amalgavirus that evolved independently for a long time.

Abundance and distribution of MjVA and MjVB in M. japonicus tissues

High-quality RNA-seq reads were mapped to MjVA and MjVB genome sequences to study the abundance and distribution of the viruses within *M. japonicus* tissue samples. For each sample, the mapped fragments were counted for MjVA or MjVB, and FPKM values were calculated for normalization (Table 1). The mapping results revealed a clear differential distribution between MjVA and MjVB.

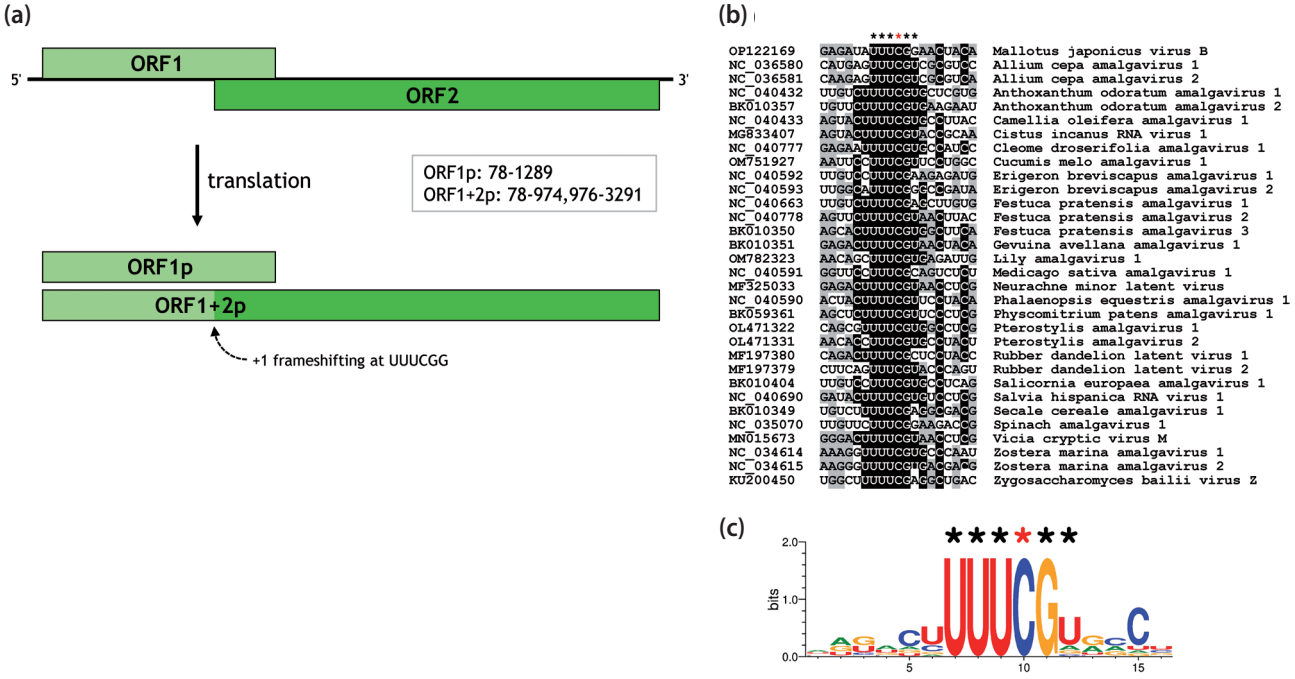

Fig. 3

Genomic organization of Mallotus japonicus virus B (MjVB)

(a) Schematic representation of the MjVB genome is depicted. ORFs and protein products are represented by shaded boxes. The genomic positions of protein-coding regions are shown in a box. **(b)** Multiple alignment of the +1 programmed ribosomal frameshifting (PRF) site sequences of MjVB and representative amalgaviruses is presented. **(c)** Sequence logo representation of the multiple alignment of the +1 PRF sites is shown. In **(b)** and **(c)**, the slippery sequence UUUCGN is marked with asterisks. The red asterisk indicates the nucleotide that is skipped by +1 PRF during translation.

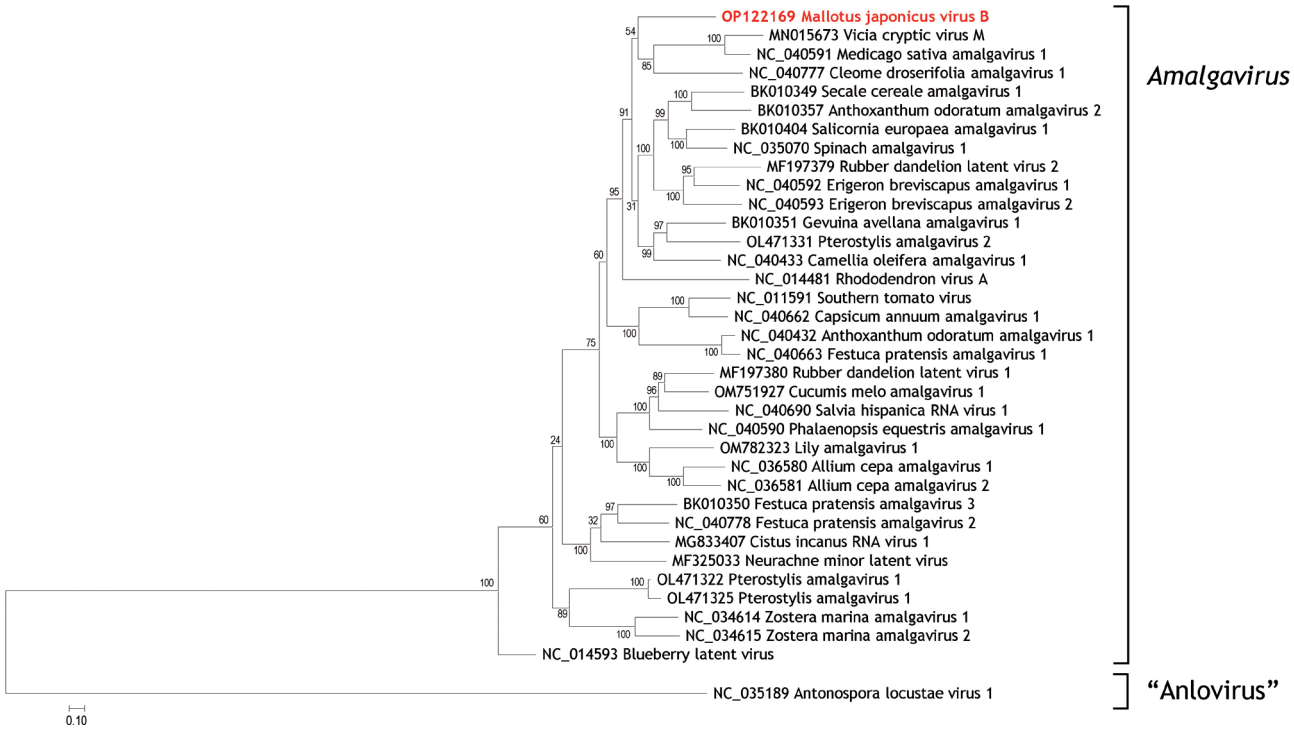

Fig. 4

Phylogenetic analysis of Mallotus japonicus virus A (MjVB)

The ORF1+2p fusion protein sequences of MjVB, 36 amalgaviruses, and Antonospora locustae virus 1 (AnloV1) were used to construct a maximum likelihood phylogenetic tree. Bootstrap values, calculated from 1000 replicates, were shown at the nodes. AnloV1, a member of the proposed genus "Anlovirus," was set as an outgroup.

MjVA was highly abundant in mature stem and bark samples with FPKM values of 1570.4 and 363.9, respectively. In case of the mature stem sample, almost 1% of all RNA-seq fragments originated from the MjVA genome. In other samples, the abundance of MjVA was relatively low (FPKM, 0.3–17.8). Conversely, the abundance of MjVB was low in all samples (FPKM, 3.1–10.7).

The differential distribution of MjVA and MjVB among *M. japonicus* samples might be due to their differences in transmission methods. Poleroviruses are restricted to and highly concentrated in the phloem tissues of host plants and transmitted by insects (mainly aphids) (Ghosh *et al.*, 2019; LaTourrette *et al.*, 2021). This finding explains the extreme abundance of MjVA in mature stems and bark,

Table 1. Virus-derived fragments in *M. japonicus* samples

Sample	SRA[a]	Total[b]	MjVA		MjVB		FPKM[d]
			mapped[c]	FPKM[d]	mapped[c]		
young leaves	SRR15027072	11262755	88	1.2	412	10.7	
mature leaves	SRR15027073	11757514	189	2.6	151	3.9	
young stems	SRR15027074	9670936	1074	17.8	169	5.0	
mature stems	SRR15027075	7260764	70964	1570.4	174	7.2	
bark	SRR15027076	7838141	17808	363.9	251	9.3	
central cylinder	SRR15027077	10221296	24	0.3	113	3.1	
inflorescence	SRR15027078	11530427	139	1.9	284	7.3	

[a]Sequence Read Archive Acc. No.; [b]Number of total high-quality fragments; [c]Number of mapped fragments; [d]Fragments per kilobase per million.

which have well-developed phloem tissues compared with that of other samples. In this study, the bark sample was obtained by dissecting a mature stem sample into the bark, which included phloem tissues, and a central cylinder, which lacked the phloem tissues. Therefore, MjVA fragments were abundant in the bark sample but scarce in the central cylinder sample. High MjVA concentrations in *M. japonicus* phloem tissues might ensure their transmission via phloem feeders, such as aphids, although no active insect-infestation was observed when the samples were collected.

Amalgaviruses are vertically transmitted from one generation to the next through seeds and generally do not develop any disease symptoms in host plants (Martin *et al.*, 2011). The probable seed-borne transmission of MjVB may ensure its systemic distribution, including inflorescences, which form gametes and eventually seeds. Therefore, the low-concentration systemic presence of MjVB may be a result of its prolonged adaptation to *M. japonicus*.

Conclusion

The high-throughput RNA-seq analysis of seven *M. japonicus* samples yielded two novel RNA viruses: polerovirus MjVA and amalgavirus MjVB. The MjVA genome encodes seven proteins from highly overlapping ORFs through various translational mechanisms, including –1 PRF, translation initiation at a non-AUG codon, and stop codon readthrough. The MjVB genome has two overlapping ORFs producing two proteins; one of which is translated by +1 PRF mechanism. Sequence comparisons and phylogenetic analyses confirmed that MjVA and MjVB are novel members of the genera *Polerovirus* and *Amalgavirus*, respectively. MjVA concentrations were higher in plant parts with well-developed phloem tissues than in other parts, while MjVB concentrations were low in all tissues. Thus, this study demonstrates the effectiveness of RNA-seq analysis for the identification of novel viruses and their distribution in plant samples.

Acknowledgments. This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Government of Korea (Grant Nos.: 2018R1A5A1025077 and 2020R1A2C1013403).

References

Arisawa M (1994): A review of the biological activity and chemistry of Mallotus japonicus (Euphorbiaceae). Phytomedicine 1, 261–269. https://doi.org/10.1016/S0944-7113(11)80074-7

Atkins JF, Loughran G, Bhatt PR, Firth AE, Baranov PV (2016): Ribosomal frameshifting and transcriptional slippage: From genetic steganography and cryptography to adventitious use. Nucleic Acids Res. 44, 7007–7078. https://doi.org/10.1093/nar/gkw530

Buchfink B, Reuter K, Drost HG (2021): Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat. Methods 18, 366–368. https://doi.org/10.1038/s41592-021-01101-x

Bushmanova E, Antipov D, Lapidus A, Prjibelski AD (2019): rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. GigaScience 8, giz100. https://doi.org/10.1093/gigascience/giz100

Capella-Gutiérrez S, Silla-Martinez JM, Gabaldon T (2009): trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Choi D, Shin C, Shirasu K, Hahn Y (2021): Two novel poty-like viruses identified from the transcriptome data of purple witchweed (Striga hermonthica). Acta Virol. 65, 365–372. https://doi.org/10.4149/av_2021_402

Choi D, Shin C, Shirasu K, Ichihashi Y, Hahn Y (2022): Artemisia capillaris nucleorhabdovirus 1, a novel member of the genus Alphanucleorhabdovirus, identified in the Artemisia capillaris transcriptome. Acta Virol. 66, 149–156. https://doi.org/10.4149/av_2022_204

Crooks GE, Hon G, Chandonia JM, Brenner SE (2004): WebLogo: a sequence logo generator. Genome Res. 14, 1188–1190. https://doi.org/10.1101/gr.849004

Csaszar K, Spackova N, Stefl R, Sponer J, Leontis NB (2001): Molecular dynamics of the frame-shifting pseudoknot from beet western yellows virus: the role of non-Watson-Crick base-pairing, ordered hydration, cation binding and base mutations on stability and unfolding. J. Mol. Biol. 313, 1073–1091. https://doi.org/10.1006/jmbi.2001.5100

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H (2021): Twelve years of SAMtools and BCFtools. GigaScience 10. https://doi.org/10.1093/gigascience/giab008

Delfosse VC, Baron MPB, Distefano AJ (2021): What we know about poleroviruses: Advances in understanding the functions of polerovirus proteins. Plant Pathol. 70, 1047–1061. https://doi.org/10.1111/ppa.13368

Depierreux D, Vong M, Nibert ML (2016): Nucleotide sequence of Zygosaccharomyces bailii virus Z: Evidence for +1 programmed ribosomal frameshifting and for assignment to family Amalgaviridae. Virus Res. 217, 115–124. https://doi.org/10.1016/j.virusres.2016.02.008

Distefano AJ, Bonacic Kresic I, Hopp HE (2010): The complete genome sequence of a virus associated with cotton blue disease, cotton leafroll dwarf virus, confirms that it is a new member of the genus Polerovirus. Arch. Virol. 155, 1849–1854. https://doi.org/10.1007/s00705-010-0764-3

Edgar RC, Taylor J, Lin V, Altman T, Barbera P, Meleshko D, Lohr D, Novakovsky G, Buchfink B, Al-Shayeb B, Banfield JF, de la Pena M, Korobeynikov A, Chikhi R, Babaian

A (2022): Petabase-scale sequence alignment catalyses viral discovery. Nature 602, 142–147. https://doi.org/10.1038/s41586-021-04332-2

Ghosh S, Kanakala S, Lebedev G, Kontsedalov S, Silverman D, Alon T, Mor N, Sela N, Luria N, Dombrovsky A, Mawassi M, Haviv S, Czosnek H, Ghanim M (2019): Transmission of a new polerovirus infecting pepper by the whitefly Bemisia tabaci. J. Virol. 93. https://doi.org/10.1128/JVI.00488-19

Goh CJ, Park D, Lee JS, Sebastiani F, Hahn Y (2018): Identification of a novel plant amalgavirus (Amalgavirus, Amalgaviridae) genome sequence in Cistus incanus. Acta Virol. 62, 122–128. https://doi.org/10.4149/av_2018_201

Hallgren J, Tsirigos KD, Pedersen MD, Almagro Armenteros JJ, Marcatili P, Nielsen H, Krogh A, Winther O (2022): DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. bioRxiv, 2022.04.08.487609. https://doi.org/10.1101/2022.04.08.487609

Haupt S, Stroganova T, Ryabov E, Kim SH, Fraser G, Duncan G, Mayo MA, Barker H, Taliansky M (2005): Nucleolar localization of potato leafroll virus capsid proteins. J. Gen. Virol. 86, 2891–2896. https://doi.org/10.1099/vir.0.81101-0

Huang LF, Naylor M, Pallett DW, Reeves J, Cooper JI, Wang H (2005): The complete genome sequence, organization and affinities of carrot red leaf virus. Arch. Virol. 150, 1845–1855. https://doi.org/10.1007/s00705-005-0537-6

Igori D, Kim SE, Kwon SY, Moon JS (2022): Complete genome sequence of Plantago asiatica virus A, a novel putative member of the genus Polerovirus. Arch. Virol. 167, 219–222. https://doi.org/10.1007/s00705-021-05265-x

Isogai M, Nakamura T, Ishii K, Watanabe M, Yamagishi N, Yoshikawa N (2011): Histochemical detection of blueberry latent virus in highbush blueberry plant. J. Gen. Plant Pathol. 77, 304–306. https://doi.org/10.1007/s10327-011-0323-0

Kim DS, Jung JY, Wang Y, Oh HJ, Choi D, Jeon CO, Hahn Y (2014): Plant RNA virus sequences identified in kimchi by microbial metatranscriptome analysis. J. Microbiol. Biotechnol. 24, 979–986. https://doi.org/10.4014/jmb.1404.04017

Knierim D, Maiss E, Menzel W, Winter S, Kenyon L (2015): Characterization of the complete genome of a novel polerovirus infecting Sauropus androgynus in Thailand. J. Phytopathol. 163, 695–702. https://doi.org/10.1111/jph.12365

Krupovic M, Dolja VV, Koonin EV (2015): Plant viruses of the Amalgaviridae family evolved via recombination between viruses with double-stranded and negative-strand RNA genomes. Biol. Direct 10, 12. https://doi.org/10.1186/s13062-015-0047-8

LaTourrette K, Holste NM, Garcia-Ruiz H (2021): Polerovirus genomic variation. Virus Evol. 7, veab102. https://doi.org/10.1093/ve/veab102

Lee JS, Goh CJ, Park D, Hahn Y (2019): Identification of a novel plant RNA virus species of the genus Amalgavirus in the family Amalgaviridae from chia (Salvia hispanica). Genes Genomics 41, 507–544. https://doi.org/10.1007/s13258-019-00782-1

Martin RR, Zhou J, Tzanetakis IE (2011): Blueberry latent virus: an amalgam of the Partitiviridae and Totiviridae. Virus Res. 155, 175–180. https://doi.org/10.1016/j.virusres.2010.09.020

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R (2020): IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. 37, 1530–1534. https://doi.org/10.1093/molbev/msaa015

Nakamura T, Yamada KD, Tomii K, Katoh K (2018): Parallelization of MAFFT for large-scale multiple sequence alignments. Bioinformatics 34, 2490–2492. https://doi.org/10.1093/bioinformatics/bty121

Nguyen Ba AN, Pogoutse A, Provart N, Moses AM (2009): NLStradamus: a simple Hidden Markov Model for nuclear localization signal prediction. BMC Bioinformatics 10, 202. https://doi.org/10.1186/1471-2105-10-202

Nibert ML, Pyle JD, Firth AE (2016): A +1 ribosomal frameshifting motif prevalent among plant amalgaviruses. Virology 498, 201–208. https://doi.org/10.1016/j.virol.2016.07.002

Park D, Goh CJ, Kim H, Hahn Y (2018): Identification of two novel amalgaviruses in the common eelgrass (Zostera marina) and in silico analysis of the amalgavirus +1 programmed ribosomal frameshifting sites. Plant Pathol. J. 34, 150–156. https://doi.org/10.5423/PPJ.NT.11.2017.0243

Park D, Hahn Y (2017): Genome sequences of spinach deltapartitivirus 1, spinach amalgavirus 1, and spinach latent virus identified in spinach transcriptome. J. Microbiol. Biotechnol. 27, 1324–1330. https://doi.org/10.4014/jmb.1703.03043

Peter KA, Liang D, Palukaitis P, Gray SM (2008): Small deletions in the potato leafroll virus readthrough protein affect particle morphology, aphid transmission, virus movement and accumulation. J. Gen. Virol. 89, 2037–2045. https://doi.org/10.1099/vir.0.83625-0

Pyle JD, Keeling PJ, Nibert ML (2017): Amalga-like virus infecting Antonospora locustae, a microsporidian pathogen of grasshoppers, plus related viruses associated with other arthropods. Virus Res. 233, 95–104. https://doi.org/10.1016/j.virusres.2017.02.015

Rai M, Rai A, Mori T, Nakabayashi R, Yamamoto M, Nakamura M, Suzuki H, Saito K, Yamazaki M (2021): Gene-metabolite network analysis revealed tissue-specific accumulation of therapeutic metabolites in Mallotus japonicus. Int. J. Mol. Sci. 22. https://doi.org/10.3390/ijms22168835

Sõmera M, Fargette D, Hébrard E, Sarmiento C, ICTV Report Consortium (2021): ICTV virus taxonomy profile: Solemoviridae 2021. J. Gen. Virol. 102. https://doi.org/10.1099/jgv.0.001707

Sato K, Kato Y, Hamada M, Akutsu T, Asai K (2011): IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. Bioinformatics 27, i85-i93. https://doi.org/10.1093/bioinformatics/btr215

Schneider TD, Stephens RM (1990): Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 18, 6097–6100. https://doi.org/10.1093/nar/18.20.6097

Shi M, Lin XD, Tian JH, Chen LJ, Chen X, Li CX, Qin XC, Li J, Cao JP, Eden JS, Buchmann J, Wang W, Xu J, Holmes EC, Zhang YZ (2016): Redefining the invertebrate RNA virosphere. Nature 540, 539–543. https://doi.org/10.1038/nature20167

Shin C, Choi D, Hahn Y (2021): Identification of the genome sequence of Zostera associated varicosavirus 1, a novel negative-sense RNA virus, in the common eelgrass (Zostera marina) transcriptome. Acta Virol. 65, 373–380. https://doi.org/10.4149/av_2021_404

Shin C, Choi D, Shirasu K, Hahn Y (2022): Identification of dicistro-like viruses in the transcriptome data of Striga asiatica and other plants. Acta Virol. 66, 157–165. https://doi.org/10.4149/av_2022_205

Smirnova E, Firth AE, Miller WA, Scheidecker D, Brault V, Reinbold C, Rakotondrafara AM, Chung BY, Ziegler-Graff V (2015): Discovery of a small non-AUG-initiated ORF in poleroviruses and luteoviruses that is required for long-distance movement. PLoS Pathog. 11, e1004868. https://doi.org/10.1371/journal.ppat.1004868

Tamura K, Stecher G, Kumar S (2021): MEGA11: Molecular Evolutionary Genetics Analysis version 11. Mol. Biol. Evol. 38, 3022–3027. https://doi.org/10.1093/molbev/msab120

Teufel F, Almagro Armenteros JJ, Johansen AR, Gislason MH, Pihl SI, Tsirigos KD, Winther O, Brunak S, von Heijne G, Nielsen H (2022): SignalP 6.0 predicts all five types of signal peptides using protein language models. Nat. Biotechnol. 40, 1023–1025. https://doi.org/10.1038/s41587-021-01156-3

Wu WP, Zhang XF, Zhu ZX, Wang HF (2021): Complete plastome sequence of Mallotus japonicus (Linn. f.) Mull. Arg. (Euphorbiaceae): a medicinal plant species endemic in East Asia. Mitochondrial DNA B Resour. 6, 1409–1410. https://doi.org/10.1080/23802359.2021.1911707

Xu Y, Ju HJ, DeBlasio S, Carino EJ, Johnson R, MacCoss MJ, Heck M, Miller WA, Gray SM (2018): A stem-loop structure in potato leafroll virus open reading frame 5 (ORF5) is essential for readthrough translation of the coat protein ORF stop codon 700 bases upstream. J. Virol. 92. https://doi.org/10.1128/JVI.01544-17

Zhang K, Xu H, Zhuang X, Zang Y, Chen J (2020): First report of vicia cryptic virus M infecting cowpea (Vigna unguiculata) in China. Plant Dis. https://doi.org/10.1094/PDIS-05-20-1148-PDN