

A novel endogenous badnavirus exists in *Alhagi sparsifolia*^{*#}

Yong-chao LI¹, Jian-guo SHEN², Guo-huan ZHAO³, Qin YAO^{†‡1}, Wei-min LI^{†‡4}

¹Institute of Life Sciences, Jiangsu University, Zhenjiang 212013, China

²Inspection & Quarantine Technology Center, Fujian Entry-Exit Inspection and Quarantine Bureau, Fuzhou 350003, China

³Key Laboratory of Agro-Biodiversity and Pest Management of Education Ministry of China, Yunnan Agricultural University, Kunming 650201, China

⁴Biotechnology Research Institute, Chinese Academy of Agricultural Sciences, Beijing 100081, China

[†]E-mail: yaoqin@ujs.edu.cn; liweimin01@caas.cn

Received Mar. 28, 2017; Revision accepted May 11, 2017; Crosschecked Mar. 12, 2018

Abstract: We report the recovery of a 7068-nt viral sequence from the “viral fossils” embedded in the genome of *Alhagi sparsifolia*, a typical desert plant. Although the full viral genome remains to be completed, the putative genome structure, the deduced amino acids and phylogenetic analysis unambiguously demonstrate that this viral sequence represents a novel species of the genus *Badnavirus*. The putative virus is tentatively termed *Alhagi bacilliform virus* (ABV). Southern blotting and inverse polymerase chain reaction (PCR) data indicate that the ABV-related sequence is integrated into the *A. sparsifolia* genome, and probably does not give rise to functional episomal virus. Molecular evidence that the ABV sequence exists widely in *A. sparsifolia* is also presented. To our knowledge, this is the first endogenous badnavirus identified from plants in the Gobi desert, and may provide new clues on the evolution, geographical distribution as well as the host range of the badnaviruses.

Key words: *Badnavirus*; Endogenous *Alhagi bacilliform virus*; Nuclear integration

<https://doi.org/10.1631/jzus.B1700171>

CLC number: Q939.46

1 Introduction

Badnaviruses, members of the plant virus family Caulimoviridae, possess an open-circular double-stranded DNA genome of 7–8 kb with non-enveloped bacilliform particles (Fauquet et al., 2005). So far, forty species are assigned to the *Badnavirus* genus (Bhat et al., 2016), which include *Commelina yellow mottle virus* (ComYMV) as the type species (Medberry et al., 1990). The badnaviruses are known to be transmitted through vegetative propagation and via vector insects such as mealybug and aphid (Harper


et al., 2004; Geering et al., 2005a; Laney et al., 2012; Seal et al., 2014; Bhat et al., 2016).

Their typical genome structure consists of three open reading frames (ORFs) solely on the plus strand (Xu et al., 2011). ORF1 and ORF2 encode proteins of unclear function(s), but previous studies have suggested that the ComYMV ORF1- and ORF2-encoded proteins are associated with the cell-wall-enriched fraction and with virions (Cheng et al., 1996), and that the *Cacao swollen shoot virus* (CSSV) ORF2 encodes a nucleic acid-binding protein (Jacquot et al., 1996). ORF3, the largest ORF, encodes a about 200-kDa polyprotein which contains conserved domains of movement protein (MP), aspartic protease (AP), a cysteine-rich zinc finger-like RNA-binding region (RB), a second cysteine-rich region (2nd CR), reverse transcriptase (RT), and ribonuclease H (RNaseH) (Medberry et al., 1990; Tzafrir et al., 1997; Wang et al., 2014; Kazmi et al., 2015). The conserved domains, particularly the RT-RNaseH, are usually employed to

[‡] Corresponding authors

^{*} Project supported by the National Natural Science Foundation of China (Nos. 31370181 and 31570146) and the Fujian Natural Science Funds for Distinguished Young Scholar (No. 2014J06008), China

[#] Electronic supplementary materials: The online version of this article (<https://doi.org/10.1631/jzus.B1700171>) contains supplementary materials, which are available to authorized users

 ORCID: Yong-chao LI, <https://orcid.org/0000-0002-8778-0972>

© Zhejiang University and Springer-Verlag GmbH Germany, part of Springer Nature 2018

discriminate between viruses within the *Badnavirus* genus (King et al., 2012). The polyprotein is specifically hydrolyzed by AP, resulting in four to five mature proteins related to viral movement and replication as well as virion packaging (Medberry et al., 1990; Jacquot et al., 1996; Hohn et al., 1997; Tzafrir et al., 1997; Hany et al., 2014). Like other members of the family Caulimoviridae, which are referred to as endogenous pararetroviruses (EPRVs), many badnaviruses also embed their genomes into host nuclear DNA (Geering et al., 2001, 2005a, 2005b; Harper et al., 2002; Philippe and Marie, 2009; Chabannes et al., 2013; Iskra-Caruana et al., 2014; Seal et al., 2014). It is proposed that this nuclear integration takes place by illegitimate recombination during DNA double strand break repair in the host genome (Staginnus and Richert-Pöggeler, 2006; Laney et al., 2012). The resulting integrants usually present diverse and complex structures, possibly because sequence rearrangement such as tandem repeat, inversion, duplication, and fragmentation occurs in the integrated viral genome (Hull et al., 2000; Umber et al., 2014). The integrants derived from most of the identified badnaviruses, including *Fig badnavirus 1* (FBV-1), *Kalanchoë top-spotting virus* (KTSV), and *Dracaena mottle virus* (DrMV), have been considered to be non-infectious (Yang et al., 2005; Su et al., 2007; Laney et al., 2012; Iskra-Caruana et al., 2014). Only integrants of *Banana streak GF virus* (BSGFV), *Banana streak OL virus* (BSOLV), and *Banana streak IM virus* (BSIMV) that infect the *Musa* species have been found to be activated as episomal infectious viruses when crossing or on abiotic stresses (Gayral et al., 2008; Chabannes et al., 2013).

The hosts of the known badnaviruses range from monocots to dicots, most of which are tropical and subtropical plants, including banana, potato, black pepper, citrus, cocoa, sugarcane, taro, and yam, as well as a few temperate crops such as red raspberry, gooseberry, and ornamental spiraea (Harper et al., 2002; Hansen et al., 2005; Staginnus and Richert-Pöggeler, 2006; Bhat et al., 2016). Therefore, the majority of the badnaviruses are distributed in tropical and subtropical climate zones, with a few in some temperate regions (Bhat et al., 2016).

Alhagi sparsifolia Shap. is a perennial subshrub belonging to the family Leguminosae, which grows in desert climates and is naturally distributed in the Gobi

desert of Northwest China and adjacent countries in Middle Asia. Recent high-resolution RNA sequencing (RNA-Seq) of the primary roots of *A. sparsifolia* (Wu et al., 2015) suggested the presence of a badnavirus. We initiated our study to determine the genome of this badnavirus, tentatively named *Alhagi bacilliform virus* (ABV), and have provided evidence that it is novel and is integrated in the *A. sparsifolia* genome. To our knowledge, ABV is the first badnavirus identified from a plant growing only in the Gobi desert, thus extending the geographical distribution of the badnaviruses.

2 Materials and methods

2.1 Plant materials

A. sparsifolia seeds collected from Taklamakan of the Xinjiang Uygur Autonomous Region, Northwest China and previously subjected to RNA-seq analyses (Wu et al., 2015) were used to clone the putative ABV sequence. To explore the distribution of the putative badnavirus in *A. sparsifolia*, more seeds were collected individually from 11 different places in Northwest China: ten from the Xinjiang Uygur Autonomous Region (Shihezi, Hutubi, Luntai, Manas Lake, Kuytun, Cele, Wensu, Yopurga, Wushi, and Alal) and one from Gansu Province (Minqin). As described by Wu et al. (2015), the seeds were first treated with concentrated sulfuric acid, and were then cultured in sterilized dishes with a layer of fully wetted filter paper in the dark at 25 °C. After 7 d, the seedlings were harvested and stored at -80 °C until use.

2.2 Total DNA extraction

By using the cetyltrimethylammonium bromide (CTAB) method (Gawel and Jarret, 1991), the total DNA of *A. sparsifolia* was extracted from 100 mg of the 7-d old seedlings, and was subsequently treated with RNase A (1 mg/ml) at 65 °C for 60 min. Following two extractions with chloroform-isoamyl (24:1, v/v), the DNA was precipitated and resuspended in a final volume of 40 µl sterile distilled deionized H₂O and stored at -20 °C until use.

2.3 Genome cloning and assembly

Nine primer sets (Table S1) were designed to generate polymerase chain reaction (PCR) products

covering the entire putative genome of ABV following a genome walking strategy. PCR was performed using the total DNA prepared from *A. sparsifolia* seedlings from Taklamakan as a template. At least five independent clones were sequenced for each amplified product. A continuous nucleotide sequence (Fig. 1a; GenBank accession KY034642) was assembled with the overlapping clones (Figs. 1b and 1c) using DNAMAN Version 5.2.2 (Lynnon Biosoft, Montreal, QC, Canada). ORF finder (<http://www.ncbi.nlm.nih.gov/projects/gorf>) was used to identify the putative ORFs in the assembled viral genome. The theoretical molecular weights (MWs) of the deduced proteins were calculated using ExPASy (http://web.expasy.org/compute_pi), and a CD-search (<http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>) was employed to identify conserved domains within the protein sequences.

2.4 Sequence analysis

Multiple sequence alignments were performed on nucleotide (nt) or deduced amino acid (aa) sequences of the putative ABV and those of the representative members in the genus *Badnavirus* using DNAMAN Version 5.2.2 (Lynnon Biosoft, Montreal, QC, Canada). Phylogenetic analysis was inferred on the 579-nt RT-RNaseH-coding region, the 456-nt AP-coding region, and the 816-nt coat protein (CP)-coding region of ABV, and 40 known badnaviruses, whose names and sequence accession numbers are listed in Table S2. *Strawberry vein banding virus* (SVBV) of the genus *Caulimovirus* was used as an outgroup since it is one of the most closely related non-badnavirus members of the family Caulimoviridae. The phylogenetic tree was built with MEGA6 using the neighbor joining method (Tamura et al., 2011). The significance of branching order was assessed by bootstrap resampling of 1000 replicates, and the cut-off value was 50%.

2.5 Southern blotting analysis

Total DNA (30 µg) prepared from the 7-d old seedlings was fully digested overnight with 100 U of *Xba*I or *Hind*III, which does not cut the sequence of the RT-RNaseH domain within ABV. The digested DNA samples, along with the equivalent amount of untreated total DNA, were separated in a 0.8% (0.008 g/ml) agarose gel, and immediately transferred to the Hybond N⁺ membrane (GE Healthcare, Life Sciences, Indianapolis, USA). Following prehybridization, the

membrane was hybridized overnight at 65 °C with the ³²P-labeled probes prepared from the 579-nt RT-RNaseH-coding region of ABV using the Prime-a-Gene[®] kit (Promega, Madison, WI, USA). The membrane was washed twice in 0.5× saline-sodium citrate (SSC) and 0.1% (1 g/L) sodium dodecyl sulfide (SDS) at 65 °C for 30 min, and exposed to X-ray film at -80 °C. Developing and fixing were performed after 2–3 d.

2.6 Inverse PCR

Total DNA (2 µg) prepared from the 7-d-old seedlings was completely digested overnight with 20 U of *Xba*I. After chloroform extraction and ethanol precipitation, 1 µg of the *Xba*I-digested DNA was ligated at 16 °C in a volume of 500 µl containing 10 µl T₄ DNA ligase overnight. The resulting ligation product was further precipitated with ethanol, and was resuspended in 10 µl of sterile distilled deionized H₂O. One microliter of the purified ligation product was immediately used as a template to perform the 1st round of PCR in a 50-µl reaction volume with the primer pair IR1-F/IR1-R. The 2nd round of PCR was conducted with 1 µl of a 200-fold dilution of the primary PCR product with the nested primer pair IR2-F/IR2-R (Table S1). The PCR conditions were as follows: initial denaturation at 95 °C for 5 min, 35 cycles at 95 °C for 30 s, 52 °C for 30 s, and 72 °C for 3 min, and final extension at 72 °C for 10 min. PCR products were analyzed by electrophoresis in 1% (0.01 g/ml) agarose gel, and a band with the expected size of approximately 3 kb was purified and ligated into the pMD18-T Vector (TaKaRa, Dalian, China) for sequencing.

2.7 Amplification of the nucleotide sequence of the RT-RNaseH domain

With a degenerate primer pair Badna-FP/Badna-RP, as show in Table S1 (Yang et al., 2003), the PCR amplification was performed in a 50-µl reaction volume containing LA Taq DNA polymerase (TaKaRa) as well as a 50-ng DNA template prepared from *A. sparsifolia* seedlings from Shihezi, Hutubi, Luntai, Manas Lake, Kuytun, Cele, Wensu, Yopurga, Wushi, Alal, and Minqin. The PCR conditions were as follows: initial denaturation at 95 °C for 5 min, 35 cycles at 95 °C for 30 s, 50 °C for 30 s, and 72 °C for 40 s, and final extension at 72 °C for 10 min. PCR products

were analyzed by 1% agarose gel electrophoresis, and a band with the expected size of approximately 580 bp was purified and ligated into the pMD18-T Vector (TaKaRa). After transformation, the positive clones were screened for sequencing.

3 Results and discussion

3.1 A putative badnavirus found in *A. sparsifolia*

Using RNA-Seq, we recently constructed a transcriptome database containing 33255 unigenes with the polyadenylated RNAs purified from the primary roots of *A. sparsifolia* from Taklamakan (Wu et al., 2015). By BlastX searching against the NCBI nr database, two unigenes, comp6965_c1 with a 246-bp length and comp22399_c0 with a size of 10144 bp, were identified to have homology with the genome of *Pagoda yellow mosaic associated virus* (PYMAV; GenBank accession KJ013302.1), a known badnavirus. The comp6965_c1 matches with nt 3526 to 3771 of PYMAV with an identity of approximately 63%. For comp22399_c0, the fragment ranging from nt 1 to 1136 matches with nt 847 to 2004 of PYMAV genome with an identity of approximately 50%, and the complementary sequence of nt 1950 to 4547 matches with nt 4224 to 6903 of PYMAV genome with an identity of 55%. The three badnavirus-like sequences were successfully cloned through PCR amplification with the corresponding primer pairs (Table S1), and were individually termed as B1, B2, and B3 (Fig. 1b). It is worth emphasizing that total DNA of *A. sparsifolia* from Taklamakan was used as the PCR template. These data collectively implied that the *A. sparsifolia* plant might harbor a badnavirus, which was tentatively named as ABV.

In order to fill gaps between the three known sequences of ABV, five sets of primer pairs (Table S1) were further designed and used to perform PCR amplification. Due to the circular genome of the badnaviruses, the primer pair of A1-F/A1-R was used to amplify the viral sequence harboring the intergenic region (IR). The obtained overlapping PCR products, along with nucleotide sequences B1, B2, and B3, were assembled, finally resulting in a continuous nucleotide sequence (7068 nt) with a G+C content of 46.9% (Fig. 1).

The IR regions of the badnaviruses are known to have a putative plant initiator methionine transfer

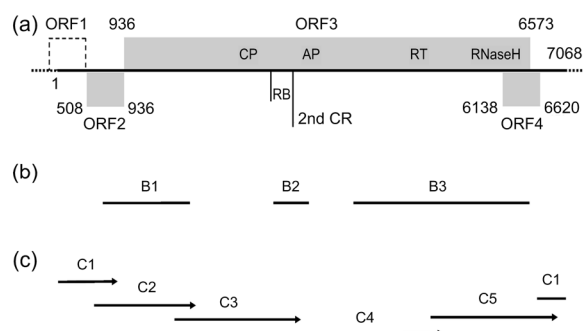


Fig. 1 Schematic genome organization of *Alhagi bacilliform virus* (ABV)

(a) The putative ORFs are indicated by open boxes. The predicted domains, CP, RB, 2nd CR, AP, RT, and RNaseH, within ORF3 are shown. The dashed lines at the 5' and 3' ends of the nucleotide sequence represent the unidentified sequence of the ABV genome. (b) Relative positions of the three badnavirus sequences derived from RNA-Seq. (c) Relative positions of cDNA clones used to assemble the ABV sequence

RNA (tRNA^{met})-binding site, the first nucleotide of which is usually used to start numbering the viral genome (Geijskes et al., 2002; Yang et al., 2003; Seal and Muller, 2007; Su et al., 2007; Kazmi et al., 2015). Because the initial 12 nucleotides (5'-TGGTATCA GAGC-3') of the putative tRNA^{met} -binding site are conserved in plant pararetroviruses, we analyzed the 7068-nt sequence by first searching for the 12 conserved nucleotides from the IR region of the assembled ABV sequence. Unfortunately, no nucleotide representing the 12-nt conserved sequence was detected, indicating that the 7068-nt ABV sequence may represent an incomplete viral genome. We, therefore, tentatively numbered the start of the ABV sequence by comparing it with the putative IR of ABV and the known badnaviruses. This needs to be refined.

Typically, the badnavirus genome encompasses three ORFs (ORF1, ORF2, and ORF3) on the plus strand (Bhat et al., 2016). Further analysis of the coding capacity of the ABV sequence identified no ORF1, probably due to the incomplete genome (Fig. 1a). However, the fragment ranging from nt 508 to 936 should be of the ABV ORF2 (Fig. 1a). The deduced protein comprises 142 aa with an MW of 16.2 kDa, and has the highest similarity at 34.9% with the putative protein encoded by the PYMAV ORF2. The fragment from nt 936 to 6573 was thought to represent the ABV ORF3, although its coding capacity was interrupted by multiple internal stop codons. The

amino acids deduced from this fragment contain domains of CP, AP, RT, RNaseH as well as two “Cys” motifs, RB and 2nd CR (Figs. 1a and 2), which are conserved in the polyproteins encoded by the ORF3 of the known badnaviruses (Medberry et al., 1990; Tzafrir et al., 1997; Wang et al., 2014; Kazmi et al., 2015). In addition, the nt 6138–6620 of ABV potentially encoded for a 160-aa protein with an MW of 17.8 kDa (Fig. 1a), and was assumed as the counterpart of the PYMAV ORF4 (Wang et al., 2014), because BLASTP analysis showed that this putative protein only produces significant alignment with the PYMAV ORF4-encoded protein.

Despite the absence of the putative tRNA^{met}-binding site and ORF1, the remaining genetic properties of this incomplete ABV genome demonstrated conclusively that ABV is indeed a badnavirus. According to the species demarcation criteria of the International Committee on Taxonomy of Viruses (ICTV) (King et al., 2012), badnaviruses with less than 80% nucleotide sequence identity or 89% amino acid sequence identity in the RT-RNaseH-coding region are regarded as species. In the current study, comparison of the putative RT-RNaseH of ABV with the same domain of the known badnaviruses exhibited amino acid sequence identities ranging from 57% to 72% and nucleotide sequence identities from 58% to 65% (Table 1), suggesting that ABV is a novel member of the genus *Badnavirus*.

3.2 Phylogenetic analysis of the conserved domains within the putative polyprotein of ABV

The relationship between the putative ABV with the other members of the genus *Badnavirus* and SVBV, a member of the genus *Caulimovirus*, was further estimated using nucleotide sequences of RT-RNaseH, CP, and AP for phylogenetic analyses. The inferred phylogenetic trees disclosed that ABV was always clustered with the badnaviruses, supporting placement of this virus within the genus *Badnavirus*. The RT-RNaseH domains of ABV and PYMAV cluster together as part of a larger cluster including the RT-RNaseH domains of nine distinct badnaviruses, BbVF, PVBV, DrMV, YNMoV, BsCVBV, GVCV, RYNV, TaBV, and GVBaV, as do the CP of ABV and PYMAV (Figs. 3a and 3b). A large cluster was also viewed in the AP-inferred phylogenetic tree, the only difference being that the AP of CyNLV, rather than TaBV, was included in this cluster (Fig. 3c). These data suggest the close relationship of ABV with the badnaviruses, in particular, PYMAV, the host of which also belongs to the family Leguminosae (Wang et al., 2014).

3.3 Integration of the ABV sequence into the *A. sparsifolia* genome

Many, if not all, badnaviruses are known to have integrants in their respective host plant genomes (Geering et al., 2001, 2005a, 2005b; Harper et al.,

Table 1 Amino acid sequence and nucleotide sequence identities of the putative RT-RNaseH domain of ABV and those of representative viruses in the genus *Badnavirus*^a

Virus	Amino acid sequence identity (%) / nucleotide sequence identity (%)											
	ABV	PYMAV	DrMV	SCBIMV	ComYMV	HBV	DBV	TaBV	KTSV	BSGFV	CSSV	
ABV	—											
PYMAV	72/65	—										
DrMV	62/62	60/63	—									
SCBIMV	59/58	60/60	63/59	—								
ComYMV	57/58	61/60	61/60	61/64	—							
HBV	63/63	62/64	64/60	60/59	64/64	—						
DBV	58/62	66/65	63/63	64/63	65/63	69/71	—					
TaBV	61/62	63/63	64/61	62/62	64/60	63/62	66/61	—				
KTSV	62/61	61/63	64/61	60/60	59/63	63/64	65/66	64/59	—			
BSGFV	60/60	64/59	63/62	60/59	59/64	65/65	63/64	61/61	67/64	—		
CSSV	60/61	60/62	60/60	60/61	60/62	67/63	62/64	64/59	57/62	55/63	—	

^a Viruses used are summarized in Table S2

Coat protein

```

ABU  LPSALQATGA +34 ENLLRETEKQFUTWR +32 PAQSTHEQDQAYADLERLQC +35 KLPQUIC +27 VLLEUCKQAARIQRSHKDLKFC +18 RKAKNWTGKPHKHTUR +12 CSCYICGDPGHFARDC
PYMAU LPSAQQTIGU +41 ENLLGETEKQIFUSWR +32 PAQSTHEQDQAYADLERLPT +35 KLPPIUG +27 VLMEUCKQAARLQRSLKDLSC +18 RAASITYGKPHKHTUR +12 CACFCICGEPDHFARDC
DBU  LPSAQQTIGA +41 ENLLGETEKLTWQWR +35 PFQSAKIQEDAYADLERISC +35 KLPDGLC +27 VLQEECKKAARSLKDLQFC +18 RKSTTYGKPHKHSUR +12 CKCFLCGEGHFARDC
TaBU  LPPAYNQAGA +41 ENLLGESEKKTWQWR +33 PYQSTAEQDQAYADLERISC +35 KMPPLIG +27 VLAEELCKKAARLQRSLKDLSC +18 RKARTYGKPHKHTUR +12 CKCFCICGEPGHFARDC
BSGFU LPSAHARQES +39 ENLLGESEKKAFHTWR +35 PKUGTTOEQDAAYKTKLSLUC +37 KLPRLHG +27 VLKDHCKEALFQSLKLNMF +22 RKNTSYGKPHKHSUR +13 CKCFACGEGHVFARDC
CSSU  LPSAQQTIGA +41 ENLLGETEKLTWQWR +35 PASGSTRIDQAYADLERLTC +35 KMPPELG +27 VLQEECKDAARFQSLKDLSC +17 RASKTYGKPHKHSUR +12 CKCYLCGEGHFARDC
KTSU  LPSAQQTIGA +39 ENLLGESEKKAFHTWR +35 PQAGTTSQDAAFKTKLSLUC +37 KLPRLG +27 VLREICQEAUFQSLKRLGFC +21 RKSTTYGKPHKHSUR +13 CKCYACGELGHFASDC
DrMU  LPSAQQLRGA +41 ENLLGEDEKKAFHTWR +33 PSQSTEEQDQAYADLERLSC +35 KMPPIIG +27 VLSIDCKQAARQKSLKDLSC +19 RKAKNRYGKPHKHTUR +12 CRCYICGEGHFARDC
ComYMU LPSAQKQDGA +39 EDLLGETERKIFUSWR +35 PULGQNTUQIAFRKLLKSLUC +35 KMPAIG +27 VLTEQCKEASVHSLKDLSC +20 RKATKYTGKADHHTUR +9 CKCYICGEGHFARDC
SCBIMU LPSAMATSGA +39 ENLLGETEKLMFTWR +35 PEQGTGQDQAYKTKLSLUC +37 KLPGLG +27 VLEEICTENNFKQLRSLNFC +19 RKARSYRGKPHSHSUR +18 CRCFUGCSTEHLKMDCK
PUBU  LPSAQQUHGA +41 ENLLGESEKTIQWWR +33 PYRGSTEEQSRAYDLERLUC +35 KLPPLIG +27 VLTDLCKKAARIQRLKDLSC +18 RKSQSYGKPHKHTUR +12 CKCFICGEGHFARDC
**::: : **::: ** * ::: * ::: : : : *::: * ** * ::: * ::: : : : ** *::: *::: *:::

```

Cysteine rich, zinc finger-like RNA binding domain (RB)

```

ABU  CSCYICGDPGHFARDCPR
PYMAU CACFCICGEPDHFARDCPR
DBU  CKCFLCGEGHFARDCPN
TaBU  CKCFCICGEPGHFARDCPT
BSGFU CKCFACGEGHVFARDCPT
CSSU  CKCYLCGEGHFARDCPN
KTSU  CKCYACGELGHFASDCPN
DrMU  CRCYICGEGHFARDCRN
ComYMU CKCYICGEGHFARDCRN
SCBIMU CRCFUGCSTEHLKMDCK
PUBU  CKCFICGEGHFARDCS
**::: ** ::: **

```

Second cysteine-rich region (2nd CR)

```

ABU  CSYCRNRTITGRISCPKLLSCLLC
PYMAU CQFCRNETRUTSRLCPACKLVAQLLC
DBU  CHTCRARDTQKHVRLCQCKFLUCSLC
TaBU  CHTCRARDTQKHVRLCQCKFLUCSLC
BSGFU CRDCKFARRDNRMDCSQCLTICALC
CSSU  CHFCCKPTNFKSRLHPCIKLITSCFMC
KTSU  CKGNCNVAAPKNRMDCPQCLTICALC
DrMU  CTYCKAPTSLVYRTKTLCLLLCCPYC
ComYMU CRSCCKQFLAG---UQCHCHAUVCYMC
SCBIMU CKRCKLTUSKGEYAYCKIKUGUCNDC
PUBU  CAFCCUQTNPENRMVCDICRLTACPMC
* * * * *

```

Aspartic protease

```

eABU  AGLTISSEUFE +87 GISUEVEEQIWSYK
PYMAU AILDGTATCCC +73 GLRIE-GPTITF-VK
DBU  AILDGTATCC +74 GURIE-GDTITF-VK
TaBU  AILDGTATUCC +74 GMRFE-GPHUTF-VK
BSGFU AILDGTAAICU +73 GLRIE-KGEVTF-VK
CSSU  AILDGTATCC +74 GLRIE-GHTITF-VK
KTSU  AILDGTATUCU +73 GIRIE-QGHUTF-VK
DrMU  AILDGTATSCU +75 GURFE-GTTITF-VK
ComYMU AILDGTATACL +74 GLRIE-KDITF-VK
SCBIMU ALLDGTATSC +75 GURLE-GRITVF-VK
PUBU  AILDGTATSCC +74 GURLE-GTTITF-VK
*::: ** ::: **

```

Reverse transcriptase

```

ABU  LLKIKUIRPSKSHRHTLAUWKSCT +11 GKEHNUYDVRQLNHNTHKQVSLPGINTIL +8 FSKFDLKSFGHQUAHMDESPWTAFLUPLGLVEWUHPFGLKNAIPAIFQRKMDHUFDDLSEFAUYIDNILIFSQTETEEHAKHL
PYMAU LLKIKUIRPSKSHRHTLAUWKSCT +11 GKEHNUYDVRQLNHNTHKQVSLPGINTIL +8 FSKFDLKSFGHQUAHMDESPWTAFLUPLGLVEWUHPFGLKNAIPAIFQRKMDHUFDDLSEFAUYIDNILIFSQTETEEHAKHL
DBU  LLKIKUIRPSKSHRHTLAUWKSCT +11 GKEHNUYDVRQLNHNTHKQVSLPGINTIL +8 FSKFDLKSFGHQUAHMDESPWTAFLUPLGLVEWUHPFGLKNAIPAIFQRKMDHUFDDLSEFAUYIDNILIFSQTETEEHAKHL
TaBU  LLKIKUIRPSKSHRHTLAUWKSCT +11 GKEHNUYDVRQLNHNTHKQVSLPGINTIL +8 FSKFDLKSFGHQUAHMDESPWTAFLUPLGLVEWUHPFGLKNAIPAIFQRKMDHUFDDLSEFAUYIDNILIFSQTETEEHAKHL
BSGFU LLKIKUIRPSKSHRHTLAUWKSCT +11 GKEHNUYDVRQLNHNTHKQVSLPGINTIL +8 FSKFDLKSFGHQUAHMDESPWTAFLUPLGLVEWUHPFGLKNAIPAIFQRKMDHUFDDLSEFAUYIDNILIFSQTETEEHAKHL
CSSU  LLKIKUIRPSKSHRHTLAUWKSCT +11 GKEHNUYDVRQLNHNTHKQVSLPGINTIL +8 FSKFDLKSFGHQUAHMDESPWTAFLUPLGLVEWUHPFGLKNAIPAIFQRKMDHUFDDLSEFAUYIDNILIFSQTETEEHAKHL
KTSU  LLKIKUIRPSKSHRHTLAUWKSCT +11 GKEHNUYDVRQLNHNTHKQVSLPGINTIL +8 FSKFDLKSFGHQUAHMDESPWTAFLUPLGLVEWUHPFGLKNAIPAIFQRKMDHUFDDLSEFAUYIDNILIFSQTETEEHAKHL
DrMU  LLKIKUIRPSKSHRHTLAUWKSCT +11 GKEHNUYDVRQLNHNTHKQVSLPGINTIL +8 FSKFDLKSFGHQUAHMDESPWTAFLUPLGLVEWUHPFGLKNAIPAIFQRKMDHUFDDLSEFAUYIDNILIFSQTETEEHAKHL
ComYMU LLKIKUIRPSKSHRHTLAUWKSCT +11 GKEHNUYDVRQLNHNTHKQVSLPGINTIL +8 FSKFDLKSFGHQUAHMDESPWTAFLUPLGLVEWUHPFGLKNAIPAIFQRKMDHUFDDLSEFAUYIDNILIFSQTETEEHAKHL
SCBIMU LLKIKUIRPSKSHRHTLAUWKSCT +11 GKEHNUYDVRQLNHNTHKQVSLPGINTIL +8 FSKFDLKSFGHQUAHMDESPWTAFLUPLGLVEWUHPFGLKNAIPAIFQRKMDHUFDDLSEFAUYIDNILIFSQTETEEHAKHL
PUBU  LLKIKUIRPSKSHRHTLAUWKSCT +11 GKEHNUYDVRQLNHNTHKQVSLPGINTIL +8 FSKFDLKSFGHQUAHMDESPWTAFLUPLGLVEWUHPFGLKNAIPAIFQRKMDHUFDDLSEFAUYIDNILIFSQTETEEHAKHL
** *::: **::: *::: **::: *::: *::: *::: *::: *::: *::: *::: *::: *::: *::: *::: *::: *::: *::: *::: *::: *::: *::: *::: *:::

```

```

ABU  +10 GLULSPTKIKIAUREVDFLGATL
PYMAU +10 GLULSPTKMKIATREIDFLGATI
DBU  +10 GLULSPTKMKIGTKTIEFLGAVI
TaBU  +10 GLULSPTKMKIGUQVDFLGATI
BSGFU +10 GLULSPTKMKIGUQVDFLGATI
CSSU  +10 GLULSPTKMKIAQREIEFLGTUI
KTSU  +10 GLULSPTKMKIGUREVDFLGATI
DrMU  +10 GLULSPTKMKIGUREVDFLGATI
ComYMU +10 GLULSPTKMKIGTPEIDFLGASL
SCBIMU +10 GLULSPTKMKIGUREVDFLGATI
PUBU  +10 GLULSPTKMKIAQREIEFLGATI
***::: **::: *::: **:::

```

Ribonuclease H

```

ABU  ITIESDGCMEGUGGICKWS +9 ERICAYASGKFTPEKSIDAEI +13 YLDKKEILIRTDQQAIIISF +7 KPSRURWLSFCDFINSGCIDUKFEHKGEMNSLADKLRL
PYMAU ITIESDGCMEGUGGICKWS +9 ERICAYASGKFTPEKSIDAEI +13 YLDKKEILIRTDQQAIIISF +7 KPSRURWLSFCDFINSGCIDUKFEHKGEMNSLADKLRL
DBU  ITIESDGCMEGUGGICKWS +9 ERICAYASGKFTPEKSIDAEI +13 YLDKKEILIRTDQQAIIISF +7 KPSRURWLSFCDFINSGCIDUKFEHKGEMNSLADKLRL
TaBU  ITIESDGCMEGUGGICKWS +9 ERICAYASGKFTPEKSIDAEI +13 YLDKKEILIRTDQQAIIISF +7 KPSRURWLSFCDFINSGCIDUKFEHKGEMNSLADKLRL
BSGFU ITIESDGCMEGUGGICKWS +9 ERICAYASGKFTPEKSIDAEI +13 YLDKKEILIRTDQQAIIISF +7 KPSRURWLSFCDFINSGCIDUKFEHKGEMNSLADKLRL
CSSU  ITIESDGCMEGUGGICKWS +9 ERICAYASGKFTPEKSIDAEI +13 YLDKKEILIRTDQQAIIISF +7 KPSRURWLSFCDFINSGCIDUKFEHKGEMNSLADKLRL
KTSU  ITIESDGCMEGUGGICKWS +9 ERICAYASGKFTPEKSIDAEI +13 YLDKKEILIRTDQQAIIISF +7 KPSRURWLSFCDFINSGCIDUKFEHKGEMNSLADKLRL
DrMU  ITIESDGCMEGUGGICKWS +9 ERICAYASGKFTPEKSIDAEI +13 YLDKKEILIRTDQQAIIISF +7 KPSRURWLSFCDFINSGCIDUKFEHKGEMNSLADKLRL
ComYMU ITIESDGCMEGUGGICKWS +9 ERICAYASGKFTPEKSIDAEI +13 YLDKKEILIRTDQQAIIISF +7 KPSRURWLSFCDFINSGCIDUKFEHKGEMNSLADKLRL
SCBIMU ITIESDGCMEGUGGICKWS +9 ERICAYASGKFTPEKSIDAEI +13 YLDKKEILIRTDQQAIIISF +7 KPSRURWLSFCDFINSGCIDUKFEHKGEMNSLADKLRL
PUBU  ITIESDGCMEGUGGICKWS +9 ERICAYASGKFTPEKSIDAEI +13 YLDKKEILIRTDQQAIIISF +7 KPSRURWLSFCDFINSGCIDUKFEHKGEMNSLADKLRL
*** *::: **::: *::: **::: *::: *::: *::: *::: *::: *::: *::: *::: *::: *::: *::: *::: *::: *::: *:::

```

Fig. 2 Comparison of amino acid sequences of the conserved domains in the putative ORF3-encoded polyprotein of ABV with those of representative badnaviruses

Virus names are showed before each sequence. Identical (*) and conserved (:) amino acids are indicated

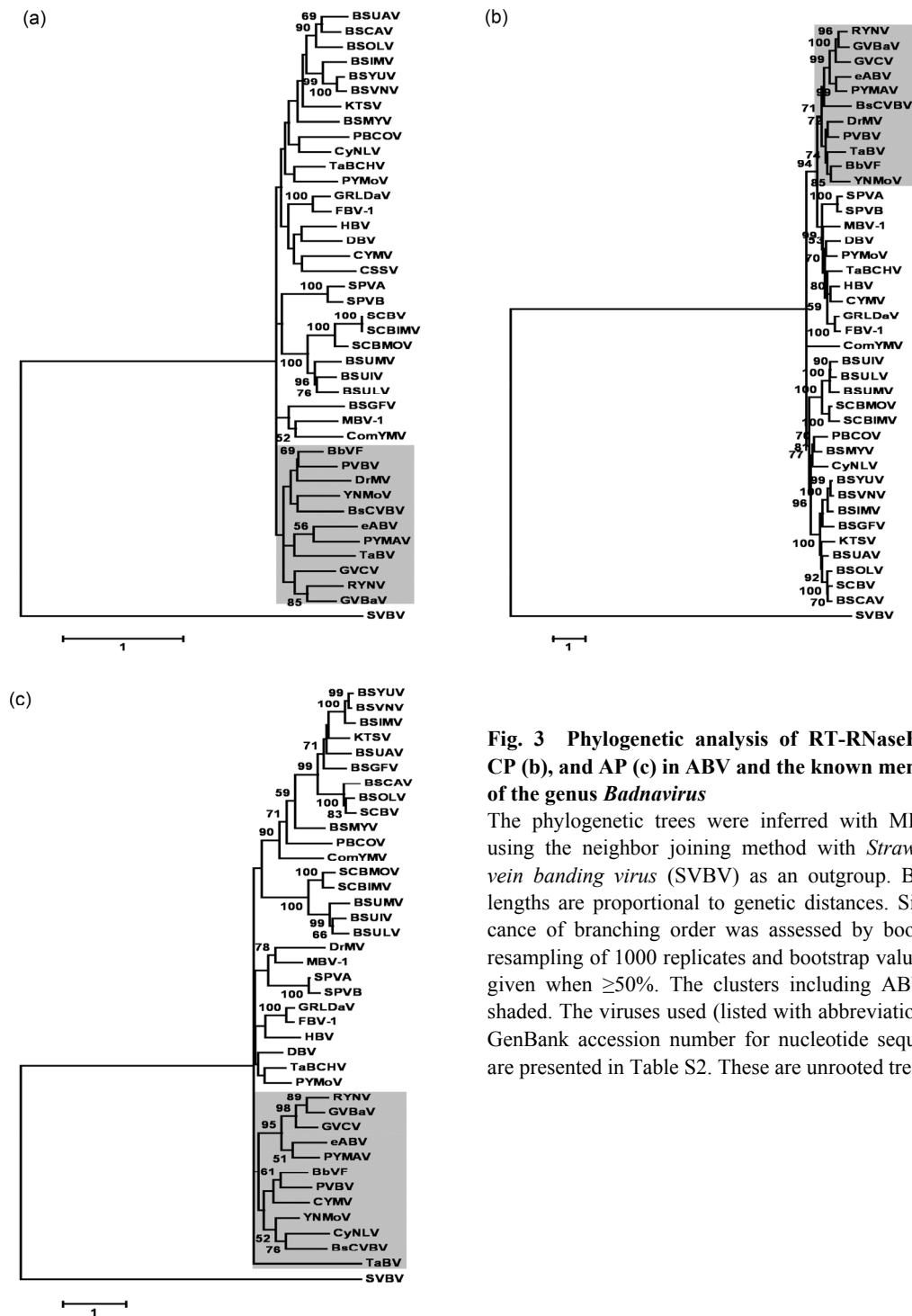


Fig. 3 Phylogenetic analysis of RT-RNaseH (a), CP (b), and AP (c) in ABV and the known members of the genus *Badnavirus*

The phylogenetic trees were inferred with MEGA6 using the neighbor joining method with *Strawberry vein banding virus* (SVBV) as an outgroup. Branch lengths are proportional to genetic distances. Significance of branching order was assessed by bootstrap resampling of 1000 replicates and bootstrap values are given when $\geq 50\%$. The clusters including ABV are shaded. The viruses used (listed with abbreviation and GenBank accession number for nucleotide sequence) are presented in Table S2. These are unrooted trees

2002; Philippe and Marie, 2009; Chabannes et al., 2013; Iskra-Caruana et al., 2014; Seal et al., 2014). To test whether ABV is endogenous to the *A. sparsifolia* genome, Southern blotting was performed to analyze the untreated total DNA from Taklamakan *A. sparsifolia* with the ^{32}P -labeled probes corresponding to the

RT-RNaseH-coding region of ABV. As shown in Fig. 4a, the undigested total DNA sample yielded a unique hybridized band at the location of the *A. sparsifolia* genomic DNA rather than at the location (about 8 kb) expected for the ABV genome, suggesting nuclear integration of the ABV sequence. One

might argue that no hybridized band corresponding to the unit length of the ABV genome was detected because there is too little, if any, episomal virus in the plant to be detected. However, considering the seriously interrupted coding capacity of the ABV ORF3 (Fig. 1), the endogenous ABV sequences should be incapable of giving rise to a functional episomal virus, but may present as long-lasting imprints within the *A. sparsifolia* genome, known as “viral fossils” (Feschotte and Gilbert, 2012). In accordance with this hypothesis, no product was generated with rolling circle amplification (RCA) to test the total DNA prepared from *A. sparsifolia* of Taklamakan (data not shown). Therefore, the recovered 7068-nt ABV sequence may reflect an ancient badnavirus that was once active but now silenced in *A. sparsifolia*.

In addition, the same probes were also used to analyze the restriction endonuclease-digested total DNA of *A. sparsifolia*. The resulting data disclosed that, for the *Hind*III-digested DNA, only a fragment of about 4.5 kb was detected (Fig. 4a, Lane 2). However, the *Xba*I-digested DNA gave rise to two discrete bands at approximately 6.5 and 3.0 kb (Fig. 4a, Lane 3), indicating at least two copies of the ABV elements in the *A. sparsifolia* genome. To further verify the endogenous nature of the ABV-related elements, we further designed two nested primer pairs according to the 579-nt RT-RNaseH-coding region of ABV, which was used to prepare the labeled probes for Southern blotting as described above, and employed the approach of inverse PCR to clone the DNA fragments representing the lower hybridization band from the *Xba*I-digested DNA. As expected, a 2925-bp DNA fragment (GenBank accession KY677913) was obtained, and sequence analysis (Fig. 4b) showed that this DNA fragment comprises a viral element corresponding to nt 5052 to 5946 of the putative ABV, which is flanked by the plant genome sequences “a” and “b” sharing homology with nt 3220 to 3480 and nt 2295 to 3173 of *Medicago truncatula* chromosome 8 clone mth2-6f11 (GenBank accession AC151460), respectively. This DNA fragment was further confirmed by a PCR analysis across the virus-plant junctions with a primer pair VPJ-F/VPJ-R (Table S1) anchored to the plant genome sequences “a” and “b”.

Taken together, these data provide additional evidence for the integration of the ABV sequence into the *A. sparsifolia* genome, and indicate that the

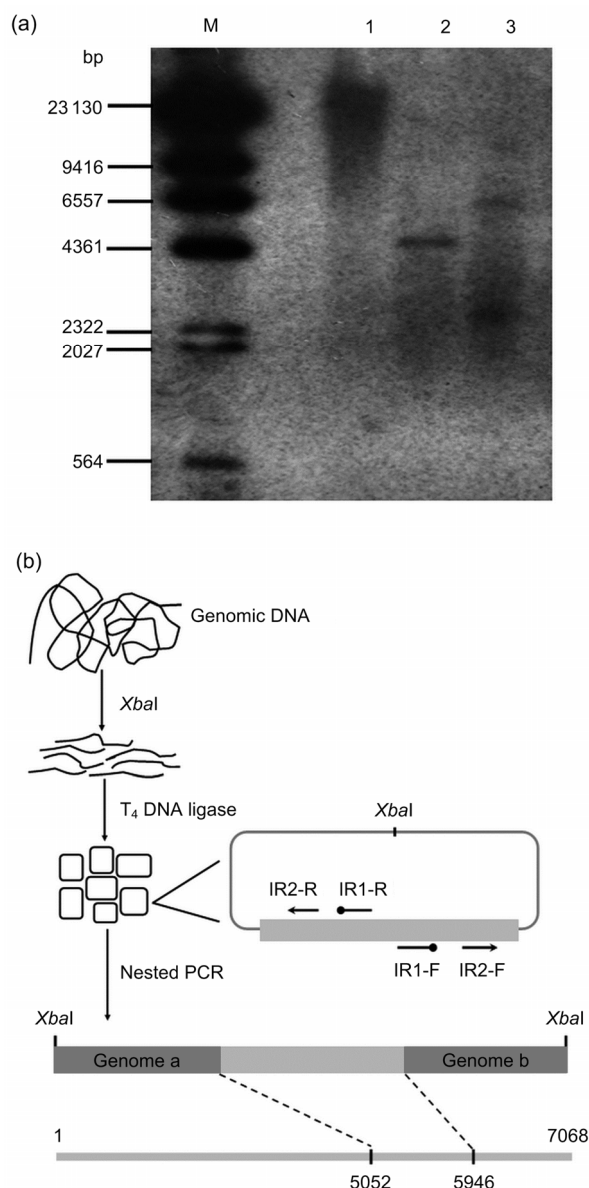


Fig. 4 Integration of the ABV-related sequences in the *A. sparsifolia* genome

(a) Southern blotting analysis of the *A. sparsifolia* genomic DNA with radiolabelled probes corresponding to the RT-RNaseH-coding region of ABV. M: λ -*Hind*III digest DNA marker; Lane 1: undigested DNA; Lane 2: DNA digested with *Hind*III; Lane 3: DNA digested with *Xba*I. (b) A schematic diagram of inverse PCR employed to clone an endogenous ABV element. The primer pairs IR1-F/IR1-R and IR2-F/IR2-R (Table S1) were designed based on nt 5203–5781 (the putative RT-RNaseH-coding region) of ABV. A 2925-bp endogenous ABV element was cloned, its position in the putative ABV genome is shown, and the grey bars “a” and “b” represent the right and left flanking plant genome sequences, respectively

integrated ABV sequence underwent complex sequence rearrangement, like integrants of other badnaviruses (Geering et al., 2001; Gregor et al., 2004; Umber et al., 2014). In accordance with this hypothesis, one DNA fragment encompassing nt 1–79, 712–795, 904–2283, and 6653–7055 of the 7059-nt ABV sequence was cloned using the primer pair of A1-F/A1-R (data not shown).

3.4 Widespread presence of the ABV sequence in *A. sparsifolia*

The presence of ABV in *A. sparsifolia* from Taklamakan prompted us to explore the distribution of the endogenous virus in this plant species. We subjected *A. sparsifolia* seedlings from 11 different places in Northwest China, Shihezi, Hutubi, Luntai, Manas Lake, Kuytun, Cele, Wensu, Yopurga, Wushi, Alal, and Minqin, to PCR detection using the degenerate primer pair Badna-FP and Badna-RP (Yang et al., 2003). Except for seedlings from three adjacent places (Wensu, Wushi, and Alal), the RT-RNaseH-coding region of ABV was amplified from all tested samples, showing widespread presence of this badnavirus in *A. sparsifolia* (Fig. 5). We could not rule out the possibility that the ABV-related sequence is embedded in the genome from plants in Wensu, Wushi, and Alal, but that we were unable to detect it because of sequence rearrangement within the entire RT-RNaseH-coding region.

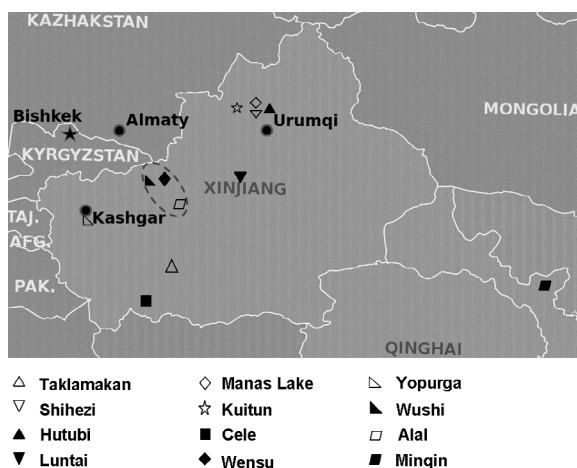


Fig. 5 Widespread presence of ABV in *A. sparsifolia*

The RT-RNaseH-coding region of ABV was detected in *A. sparsifolia* seedlings from nine different places of the Xinjiang Uygur Autonomous Region and Gansu Province, China, except for those from Wensu, Wushi, and Alal, the three adjacent places which are circled with dashed lines

4 Conclusions

This study reports evidence for an ancient endogenous badnavirus, tentatively termed ABV, in *A. sparsifolia*, a typical desert plant which differs from the known badnavirus hosts which are tropical, subtropical, and temperate plants. While the full genome sequence of ABV remains to be completed, the 7068-nt viral sequence recovered from the endogenous badnavirus elements in *A. sparsifolia* may shed fresh light on the evolution, geographical distribution as well as the host range of the badnaviruses. Dating of the endogenization event of ABV may not be possible. However, it seems to have occurred after the speciation of *A. sparsifolia* because no ABV sequence has so far been identified from 15 *Leguminosae* species with complete genomes available in GenBank. The endogenous badnavirus sequences in *A. sparsifolia* imply a potential contribution to the complexity and evolution of the host genome, and merit further investigation.

Acknowledgements

We thank Dr. Wang-bin ZHANG (College of Plant Science, Tarim University, the Xinjiang Uygur Autonomous Region, China) and Dr. Xin-wu PEI (Biotechnology Research Institute, Chinese Academy of Agricultural Sciences, Beijing, China) for helping collect the seeds of *A. sparsifolia*, and Dr. Hong-yan SHAN (Institute of Botany, Chinese Academy of Sciences, Beijing, China) for her helpful suggestion to construct the phylogenetic trees.

Compliance with ethics guidelines

Yong-chao LI, Jian-guo SHEN, Guo-huan ZHAO, Qin YAO, and Wei-min LI declare that they have no conflict of interest.

This article does not contain any studies with human or animal subjects performed by any of the authors.

References

- Bhat AI, Hohn T, Selvarajan R, 2016. Badnaviruses: the current global scenario. *Viruses*, 8(6):177. <https://doi.org/10.3390/v8060177>
- Chabannes M, Baurens FC, Duroy PO, et al., 2013. Three infectious viral species lying in wait in the banana genome. *J Virol*, 87(15):8624-8637. <https://doi.org/10.1128/JVI.00899-13>
- Cheng CP, Lockhart BEL, Olszewski NE, 1996. The ORF I and II proteins of *Commelina* yellow mottle virus are virion-associated. *Virology*, 223(2):263-271. <https://doi.org/10.1006/viro.1996.0478>
- Fauquet CM, Mayo MA, Maniloff J, et al., 2005. Virus

- taxonomy, classification and nomenclature of viruses. Eighth Report of the International Committee on the Taxonomy of Viruses. Elsevier Academic Press, San Diego.
- Feschotte C, Gilbert C, 2012. Endogenous viruses: insights into viral evolution and impact on host biology. *Nat Rev Genet*, 13(4):283-296.
<https://doi.org/10.1038/nrg3199>
- Gawel NJ, Jarret RL, 1991. A modified CTAB DNA extraction procedure for *Musa* and *Ipomoea*. *Plant Mol Biol Rep*, 9(3):262-266.
<https://doi.org/10.1007/BF02672076>
- Gayral P, Noa-Carrazana JC, Lescot M, et al., 2008. A single *Banana streak virus* integration event in the banana genome as the origin of infectious endogenous pararetrovirus. *J Virol*, 82(13):6697-6710.
<https://doi.org/10.1128/JVI.00212-08>
- Geering ADW, Olszewski NE, Dahal G, et al., 2001. Analysis of the distribution and structure of integrated *Banana streak virus* DNA in a range of *Musa* cultivars. *Mol Plant Pathol*, 2(4):207-213.
<https://doi.org/10.1046/j.1464-6722.2001.00071.x>
- Geering ADW, Olszewski NE, Harper G, et al., 2005a. Banana contains a diverse array of endogenous badnaviruses. *J Gen Virol*, 86(2):511-520.
<https://doi.org/10.1099/vir.0.80261-0>
- Geering ADW, Pooggin MM, Olszewski NE, et al., 2005b. Characterisation of Banana streak Mysore virus and evidence that its DNA is integrated in the B genome of cultivated *Musa*. *Arch Virol*, 150(4):787-796.
<https://doi.org/10.1007/s00705-004-0471-z>
- Geijskes RJ, Braithwaite KS, Dale JL, et al., 2002. Sequence analysis of an Australian isolate of *Sugarcane bacilliform badnavirus*. *Arch Virol*, 147(12):2393-2404.
<https://doi.org/10.1007/s00705-002-0879-2>
- Gregor W, Mette MF, Staginnus C, et al., 2004. A distinct endogenous pararetrovirus family in *Nicotiana tomentosiformis*, a diploid progenitor of polyploid tobacco. *Plant Physiol*, 134(3):1191-1199.
<https://doi.org/10.1104/pp.103.031112>
- Hansen CN, Harper G, Heslop-Harrison JS, 2005. Characterisation of pararetrovirus-like sequences in the genome of potato (*Solanum tuberosum*). *Cytogenet Genome Res*, 110(1-4):559-565.
<https://doi.org/10.1159/000084989>
- Hany U, Adams IP, Glover R, et al., 2014. The complete genome sequence of *Piper yellow mottle virus* (PYMoV). *Arch Virol*, 159(2):385-388.
<https://doi.org/10.1007/s00705-013-1824-2>
- Harper G, Hull R, Lockhart B, et al., 2002. Viral sequences integrated into plant genomes. *Ann Rev Phytopathol*, 40(1):119-136.
<https://doi.org/10.1146/annurev.phyto.40.120301.105642>
- Harper G, Hart D, Moul S, et al., 2004. *Banana streak virus* is very diverse in Uganda. *Virus Res*, 100(1):51-56.
<https://doi.org/10.1016/j.virusres.2003.12.024>
- Hohn T, Fütterer J, Hull R, 1997. The Proteins and functions of plant pararetroviruses: knowns and unknowns. *Crit Rev Plant Sci*, 16(1):133-161.
<https://doi.org/10.1080/713608145>
- Hull R, Harper G, Lockhart B, 2000. Viral sequences integrated into plant genomes. *Trends Plant Sci*, 5(9):362-365.
[https://doi.org/10.1016/S1360-1385\(00\)01723-4](https://doi.org/10.1016/S1360-1385(00)01723-4)
- Iskra-Caruana ML, Duroy PO, Chabannes M, et al., 2014. The common evolutionary history of badnaviruses and banana. *Infect Genet Evol*, 21:83-89.
<https://doi.org/10.1016/j.meegid.2013.10.013>
- Jacquot E, Hagen LS, Jacquemond M, et al., 1996. The open reading frame 2 product of cacao swollen shoot badnavirus is a nucleic acid-binding protein. *Virology*, 225(1):191-195.
<https://doi.org/10.1006/viro.1996.0587>
- Kazmi SA, Yang Z, Hong N, et al., 2015. Characterization by small RNA sequencing of *Taro bacilliform CH virus* (TaBCHV), a novel badnavirus. *PLoS ONE*, 10(7):e0134147.
<https://doi.org/10.1371/journal.pone.0134147>
- King AMQ, Adams MJ, Lefkowitz EJ, et al., 2012. Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses. Academic Press, San Diego, CA, USA.
- Laney AG, Hassan M, Tzanetakis IE, 2012. An integrated badnavirus is prevalent in fig germplasm. *Phytopathology*, 102(12):1182-1189.
<https://doi.org/10.1094/PHYTO-12-11-0351>
- Medberry SL, Lockhart BE, Olszewski NE, 1990. Properties of *Commelina* yellow mottle virus's complete DNA sequence, genomic discontinuities and transcript suggest that it is a pararetrovirus. *Nucleic Acids Res*, 18(18):5505-5513.
<https://doi.org/10.1093/nar/18.18.5505>
- Philippe G, Marie I, 2009. Phylogeny of *Banana streak virus* reveals recent and repetitive endogenization in the genome of its banana host (*Musa* sp.). *J Mol Evol*, 69(1):65-80.
<https://doi.org/10.1007/s00239-009-9253-2>
- Seal S, Muller E, 2007. Molecular analysis of a full-length sequence of a new yam badnavirus from *Dioscorea sansibarensis*. *Arch Virol*, 152(4):819-825.
<https://doi.org/10.1007/s00705-006-0888-7>
- Seal S, Turaki A, Muller E, et al., 2014. The prevalence of badnaviruses in West African yams (*Dioscorea cayenensis-rotundata*) and evidence of endogenous pararetrovirus sequences in their genomes. *Virus Res*, 186:144-154.
<https://doi.org/10.1016/j.virusres.2014.01.007>
- Staginnus C, Richert-Pöggeler KR, 2006. Endogenous pararetroviruses: two-faced travelers in the plant genome. *Trends Plant Sci*, 11(10):485-491.
<https://doi.org/10.1016/j.tplants.2006.08.008>
- Su L, Gao S, Huang Y, et al., 2007. Complete genomic sequence of *Dracaena* mottle virus, a distinct badnavirus. *Virus Genes*, 35(2):423-429.
<https://doi.org/10.1007/s11262-007-0102-3>

- Tamura K, Peterson D, Peterson N, et al., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol*, 28(10):2731-2739. <https://doi.org/10.1093/molbev/msr121>
- Tzafirir I, Ayala-Navarrete L, Lockhart BEL, et al., 1997. The N-terminal portion of the 216-kDa polyprotein of *Com-melina* yellow mottle badnavirus is required for virus movement but not for replication. *Virology*, 232(2): 359-368. <https://doi.org/10.1006/viro.1997.8569>
- Umber M, Filloux D, Muller E, et al., 2014. The genome of African yam (*Dioscorea cayenensis-rotundata* complex) hosts endogenous sequences from four distinct badnavirus species. *Mol Plant Pathol*, 15(8):790-801. <https://doi.org/10.1111/mpp.12137>
- Wang Y, Cheng X, Wu X, et al., 2014. Characterization of complete genome and small RNA profile of pagoda yellow mosaic associated virus, a novel badnavirus in China. *Virus Res*, 188:103-108. <https://doi.org/10.1016/j.virusres.2014.04.006>
- Wu H, Zhang Y, Zhang W, et al., 2015. Transcriptomic analysis of the primary roots of *Alhagi sparsifolia* in response to water stress. *PLoS ONE*, 10(3):e0120791. <https://doi.org/10.1371/journal.pone.0120791>
- Xu D, Mock R, Kinard G, et al., 2011. Molecular analysis of the complete genomic sequences of four isolates of *Gooseberry vein banding associated virus*. *Virus Genes*, 43(1):130-137. <https://doi.org/10.1007/s11262-011-0614-8>
- Yang IC, Hafner GJ, Dale JL, et al., 2003. Genomic characterisation of *Taro bacilliform virus*. *Arch Virol*, 148(5): 937-949. <https://doi.org/10.1007/s00705-002-0969-1>
- Yang Z, Nicolaisen M, Olszewski NE, et al., 2005. Sequencing, improved detection, and a novel form of *Kalanchoë top-spotting virus*. *Plant Dis*, 89(3):298-302. <https://doi.org/10.1094/PD-89-0298>

List of electronic supplementary materials

Table S1 Primers used in this study

Table S2 *Badnaviruses* species for phylogenetic analysis

中文概要

题 目: 疏叶骆驼刺中包含一种新杆状 DNA 病毒

目 的: 新杆状 DNA 病毒的分离与鉴定。

创新点: 首次在高寒地区代表性植物——疏叶骆驼刺中发现杆状 DNA 病毒, 为研究杆状 DNA 病毒的进化、地理分布及寄主范围提供了新证据。

方 法: 利用分段聚合酶链式反应 (PCR) 克隆疏叶骆驼刺杆状病毒 (*Alhagi bacilliform virus*, ABV) 的基因组序列; 通过基因组分析、序列比对和进化树分析阐明 ABV 的进化地位; 用 Southern 印迹杂交和反向 PCR 分析 ABV 序列与宿主基因组的关系; 并通过 PCR 检测确定 ABV 在我国西北地区疏叶骆驼刺中的分布。

结 论: 本研究获得了 7068 nt 的 ABV 基因组序列, 根据基因组结构、保守序列比对及进化树分析, 推测 ABV 是一种新杆状 DNA 病毒。分子检测证据表明, ABV 基因组序列已整合进入疏叶骆驼刺基因组中, 但没有产生游离病毒。此外, 对我国西北 11 个不同地区的疏叶骆驼刺进行 PCR 检测, 结果显示其中 9 个地区的疏叶骆驼刺均含有 ABV 序列, 由此表明 ABV 在我国西北地区的疏叶骆驼刺中广泛存在。

关键词: 杆状 DNA 病毒; 内生疏叶骆驼刺杆状病毒; 基因组整合