Arch Virol (2017) 162:1363–1371
DOI 10.1007/s00705-017-3235-2

BRIEF REPORT

Molecular characterization of previously elusive badnaviruses associated with symptomatic cacao in the New World

Nomatter Chingandu[1] · Muhammad Zia-ur-rehman[2] · Thyail N. Sreenivasan[3] · Surendra Surujdeo-Maharaj[3] · Pathmanathan Umaharan[3] · Osman A. Gutierrez[4] · Judith K. Brown[1]

Received: 4 August 2016 / Accepted: 2 January 2017 / Published online: 25 January 2017

Abstract Suspected virus-like symptoms were observed in cacao plants in Trinidad during 1943, and the viruses associated with these symptoms were designated as strains A and B of cacao Trinidad virus (CTV). However, viral etiology has not been demonstrated for either phenotype. Total DNA was isolated from symptomatic cacao leaves exhibiting the CTV A and B phenotypes and subjected to Illumina HiSeq and Sanger DNA sequencing. Based on *de novo* assembly, two apparently full-length badnavirus genomes of 7,533 and 7,454 nucleotides (nt) were associated with CTV strain A and B, respectively. The Trinidad badnaviral genomes contained four open reading frames, three of which are characteristic of other known badnaviruses, and a fourth that is present in only some badnaviruses. Both badnaviral genomes harbored hallmark caulimovirus-like features, including a $tRNA^{Met}$ priming site, a TATA box, and a polyadenylation-like signal. Pairwise comparisons of the RT-RNase H region indicated that the Trinidad isolates share 57-71% nt sequence identity with other known badnaviruses. Based on the system for badnavirus species demarcation in which viruses with less than 80% nt sequence identity in the RT-RNase gene are considered members of separate species, these isolates represent two previously unidentified badnaviruses, herein named cacao mild mosaic virus and cacao yellow vein banding virus, making them the first cacao-infecting badnaviruses identified thus far in the Western Hemisphere.

Theobroma cacao L. (cacao) is the source of cocoa beans and an economically important crop in the neotropics and in West Africa, where it was introduced and grown for commercial production over a century ago [10]. A major economic constraint to cacao production in West Africa is cacao swollen shoot virus (CSSV) (genus *Badnavirus*; family *Caulimoviridae*), causing reduced yields, tree decline, and tree death 3-5 years after infection.

Badnaviruses are pararetroviruses belonging to the family *Caulimoviridae*. They have a circular double-stranded DNA genome of 7.2-9.2 kilobase pairs (kbp) in size encapsidated in a non-enveloped virion with bacilliform morphology, 60-900 nm in length and 30 nm in diameter [16]. Badnaviral genomes characteristically encode three open reading frames (ORFs), ORFs 1-3, but they may have one or more additional ORFs. All ORFs are encoded on the viral-sense strand [16]. The function of the ORF1 protein is unknown, whereas, in CSSV, ORF2 encodes a 14-kDa protein that is involved in DNA and RNA binding [13]. ORF3 encodes a 25-kDa polyprotein, with domains encoded near the 5'-end attributed to viral movement, as well as the viral capsid, aspartic protease, viral reverse transcriptase (RT), and ribonuclease H (RNase H) proteins. In CSSV, two additional ORFs overlap with ORF3 and are referred to as ORFX and Y. They are predicted to encode proteins of 13 and 14 kDa in size,

Electronic supplementary material The online version of this article (doi:10.1007/s00705-017-3235-2) contains supplementary material, which is available to authorized users.

✉ Judith K. Brown
jkbrown@email.arizona.edu

[1] School of Plant Sciences, University of Arizona, 1140 E, South Campus Drive, Tucson, AZ 85721, USA

[2] IAGS, University of Punjab, Lahore, Pakistan

[3] Cocoa Research Centre, The University of the West Indies, St. Augustine, Trinidad and Tobago

[4] USDA-ARS Subtropical Horticultural Research Station, Miami, FL 33158, USA

respectively, but their functions are unknown [14]. Several badnaviruses have ORFs homologous to the CSSV ORFY, referred to as ORF4 or ORF6 [6].

In Trinidad, virus-like symptoms were reported in *T. cacao* trees by Posnette in 1943 [24]. Disease symptoms were manifested as 'red mottling' or 'vein-clearing' [24], and the suspected viruses were referred to as strain A and B, respectively [3, 4], of the Cacao Trinidad virus (CTV) [30], with the 'vein-clearing' phenotype later revised to 'yellow vein-banding' [17]. Cacao plants affected by strain A show feather-like red vein banding on newly developing leaves, which disappears as the leaves mature (Fig. 1a), and red discoloration of young pods in ICS 6 and 8 clones. In contrast, strain B causes persistent yellow vein-banding in major and minor veins of the mature leaves that may be accompanied by red vein-banding [3]. Asymptomatic leaves can apparently support virus infection, as has been demonstrated by positive mealybug transmission tests [26]. For virus detection, the ICS 6 clone has been used as a differential indicator for both putative viral strains. Inoculation of ICS 6 with strain A produces red vein-banding, while strain B causes vein-banding and mosaic, referred to as mild and severe symptoms, respectively [3].

Transmission studies have shown that the mealybugs *Pseudococcus citri* (Risso), *P. brevipes* (Ckll.), *P. comstocki* (Kuw.), and *Ferrisia virgata* (Cockerell) transmit both strains of CTV [4, 17]. Transmission by these four mealybug vectors requires a minimal acquisition access period of ~30 min, and an inoculation access period of ~90 min, suggesting that CTV is transmitted in a non-persistent manner by multiple vector species [17].

Symptoms of CTV were first observed in cacao trees in the Diego Martin Valley in the northwest region of the Northern Range, and the disease was first thought to be restricted to this area [24]. However, island-wide surveys revealed the presence of both strains in the Santa Cruz and Maracas [4, 24] and Blanchisseuse and Toco valleys [4, 28], suggesting that the suspected virus had been

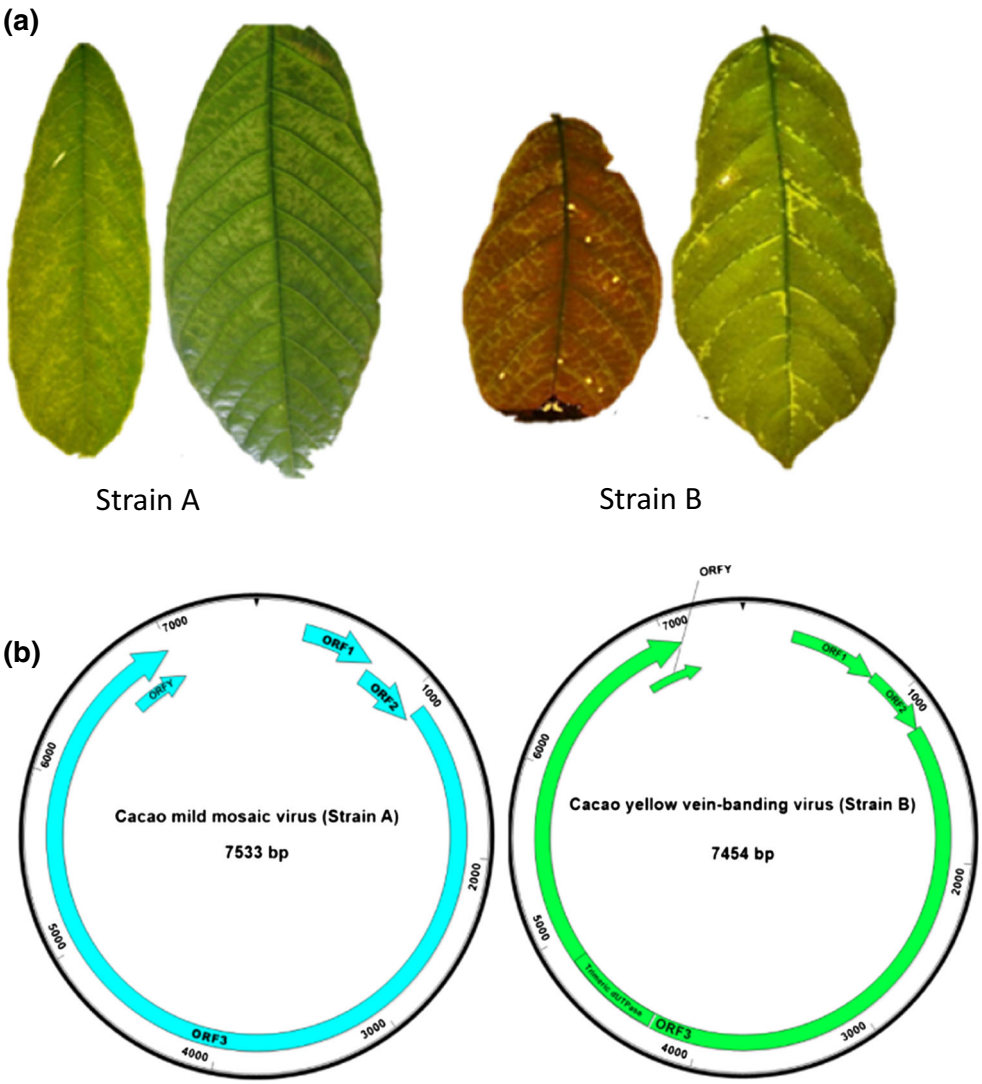

Fig. 1 Characteristic symptoms of the former cacao Trinidad virus strain A and B isolates on young and mature cacao leaves (a) and the genome map of the newly identified cacao mild mosaic virus (CaMMV) and cacao yellow vein-banding virus (CYVBV) (b). Putative coding regions for open reading frames (ORFs) 1, 2, 3 and Y, predicted using the NCBI ORF Finder tool, are indicated by filled arrows. The CaMMV and CYVBV genomes are 7,533 and 7,454 base pairs in length, respectively. The black arrowhead indicates the position of the first nucleotide and of the 5'end of the predicted $tRNA^{Met}$ binding site, which is required for priming reverse transcription of the negative strand

present in Trinidad for some time. Following a 1950s government mandated-eradication program to remove all symptomatic trees, virus-like symptoms were not observed on either island until fourteen years ago, when strain-A- and B-like symptoms were discovered in trees maintained in the International Cocoa Genebank, Trinidad (ICGT) [26].

The Cocoa Research Centre (CRC, formerly, the Cocoa Research Unit) provides cacao germplasm to many cacao-growing countries and receives and sends germplasm to the International Cocoa Quarantine Centre-Reading (ICQCR) in Reading, UK, where germplasm is quarantined and indexed by grafting to CSSV-susceptible 'Amelonado' seedlings. Grafted scions are inspected for symptom development over a period of two years before re-distribution to cocoa-growing countries. In 2002, the ICQCR reported virus-like symptoms on cacao accession ICS 76 from Trinidad following virus indexing. In 2005, virus-like symptoms were detected on ICS 76 budwood from Trinidad provided to the ICQCR [26]. Inspection of the ICS 76 mother tree at ICGT revealed foliar red mottling symptoms on flush growth, like that reported for CTV strain A [4] (Fig. 1a), and clone ICS 27 exhibited yellow vein-banding symptoms reminiscent of strain B infection, indicating that neither strain of the suspect virus had been eradicated.

Despite suspected badnavirus-like etiology, molecular detection of badnavirus in CTV-symptomatic cacao leaves has failed when using polymerase chain reaction (PCR) amplification with primers based on seven available CSSV genome sequences or when using 'universal' badnavirus primers [32], suggesting that CTV may be caused by CSSV-unrelated virus(es) and that it differs sufficiently from other known badnaviruses around which 'universal' primers were designed (authors, unpublished results). Further, no virus-like particles were observed by transmission electron microscopy (TEM) in negatively-stained leaf dip preparations of CTV strain A and strain B symptomatic leaves (P. Umaharan, unpublished results).

To determine the identity of the suspect badnavirus(es), cacao leaves exhibiting symptoms characteristic of CTV strains A and B were collected from the ICS 76 and ICS 27 cacao trees maintained at ICGT. The trees were initially established by vegetative propagation from plants maintained in collections first established during the 1930s. Budwood collected from strain A and B symptomatic ICS 76 and ICS 27 was grafted onto SCA 6 rootstock, and symptoms that developed on newly developing leaves were reminiscent of those previously described for CTV strains A and B (Fig. 1a). The leaves were washed with water, preserved in 100% glycerol, and sent to the School of Plant Sciences, University of Arizona, Tucson, AZ, USA, where they were processed for analysis.

Total DNA was isolated from 100 mg of leaf tissue, using the cetyl trimethylammonium bromide (CTAB) method [9] with modifications. To minimize potential carryover of non-viral circular DNA (e.g., plant mitochondrial and chloroplast), purified DNA was filtered through a 0.1-μm-pore Ultrafree-MC column (Millipore, MA, USA) and collected by microcentrifugation at 12,000×*g* for 4 min. The DNA was precipitated in 100% ethanol, with the addition of sodium acetate to a final concentration of 0.3 M, and the pelleted DNA was redissolved in 100 μL of low TE buffer, pH 8.0. The purified DNA was subjected to DNA sequencing using the Illumina HiSeq platform and, for follow-up validation, to PCR amplification and Sanger DNA sequencing of cloned viral fragments.

For Illumina sequencing, paired-end libraries were prepared using a TruSeq PE cluster kit, with a mean insert size of 350 bp, and individually tagged. Libraries were sequenced using the Illumina HiSeq 2500 platform at the University of Arizona Genomics Core (UAGC, Tucson, Arizona, USA). *De novo* assembly was carried out using SeqMan NGen v.12 (DNASTAR, WI, USA), using the option to filter background *T. cacao* genome sequences [2, 20] (GenBank accession numbers FR7222157.1, CM001879.1 to CM001888.1, and KE132922.1), the cacao chloroplast genome (GenBank accession no.NC014676.2), and the mitochondrial genome of a related malvaceous species, *Gossypium hirsutum* L. (GenBank accession no. JX065074.1). The assembled contigs were annotated separately using BLAST2GO software [7] and subjected to a BLASTn search in the NCBI GenBank database [1]. The two apparently full-length badnaviral genome sequences were arranged with the first nucleotide of the predicted $tRNA^{Met}$ primer binding site as coordinate number one.

The two available cacao genome sequences (GenBank accession numbers FR7222157.1, CM001879.1 to CM001888.1, and KE132922.1) were searched for possible integration of the Illumina CaMMV and CYVBV genome sequences using the Virus-host integration option in SeqMan NGen v.12 software (DNASTAR, Madison, WI).

ORFs encoding more than100 amino acids (aa) were identified in both of the apparently full-length badnaviral genomes, using the NCBI ORF Finder tool with the standard genetic code. The BLASTp search tool [1] was used to identify homologous ORFs and to compare ORFs between the previously unidentified strain A and strain B genomic sequences from cacao plants in Trinidad and the full-length badnavirus genome sequences available in GenBank. The conserved badnaviral domains were predicted using the NCBI Conserved Domain Database (CDD) tool on NCBI [18].

To analyze the assembled genome sequence, the genome sequences of 120 badnavirus isolates, representing 35

species, were downloaded from the NCBI GenBank database, and a haplotype search was performed using the FaBox tool [31] to eliminate redundant sequences. The resultant 84 haplotypes were used to determine the pairwise nt sequence identities and phylogenetic relationships in the genome segment encoding the conserved RT-RNase H region and the full-length genomic sequences. The 84 haplotypes were aligned using the MUSCLE algorithm, followed by pairwise nt sequence analysis using the Sequence Demarcation Tool (SDTv1.2) [22]. Similarly, the 84 partial and complete genome sequences were subjected to phylogenetic analysis using the maximum-likelihood (ML) algorithm implemented in MEGA 6 [29]. Trees were reconstructed using the general time-reversible model and gamma distribution with invariable sites, with 1000 bootstrap replicates and a confidence interval of >70%.

To confirm the Illumina sequences, circular viral DNA was preferentially enriched by rolling circle-amplification (RCA) using a Templiphi kit (GE Healthcare, NJ, USA), which employs phi29 DNA polymerase [8], according to the manufacturer's instructions, with, previously described modifications [25, 27]. The RCA product was used as a template for PCR amplification of badnaviral genomic fragments. Based on the two putative badnavirus genome sequences determined by Illumina sequencing, four pairs of primers, CaMMV_1F/1R and CaMMV 2F/2R for strain A, and CYVBV 1F/1R and CYVBV 2F/2R for strain B (Supplemental Table 1), were designed and used to PCR-amplify two approximately 4,000-bp fragments of each genome, with overlapping fragments of 250-500 bp.

Each PCR amplification reaction was carried out using Invitrogen CloneAmp^{TM} HiFi PCR Premix (Clonetech, CA, USA) according to the manufacturer's instructions. The reaction mixture contained 1X CloneAmp^{TM} HiFi PCR Premix, 0.2 μM each of the reverse and forward primer, 2 μL of the RCA product as template, and nuclease-free water to a total volume of 50 μL. PCR conditions were as follows: initial denaturation at 98 °C for 2 min, followed by 40 cycles of denaturation at 98 °C for 20 s, annealing at 55 °C for 15 s, and extension at 72 °C for 4 min, and a final extension at 72 °C for 10 min. The PCR products were fractionated by electrophoresis on a 0.8% agarose gel stained with GelGreen stain (10 μL/mL; Biotium, CO, USA) in 1X Tris-acetate EDTA (TAE) buffer, pH 8.0. Bands of ~4 kbp in size were excised and gel-purified using an Illustra GFX PCR DNA and Gel Band Purification kit (GE Healthcare, NJ, USA) according to the manufacturer's instructions, and the concentration of the DNA was determined using a NanoDrop 2000 UV-Vis spectrophotometer (Thermo Scientific, DE, USA). The purified DNA was ligated into the pGEM5 plasmid vector (Promega, WI, USA), which had been digested using the restriction endonuclease *Not* I (New England Biolabs, CA, USA). Ligation and transformation were carried out using an In-Fusion HD Cloning kit (Clontech, CA, USA), according to the manufacturer's instructions. Insert sizes were confirmed by isolating plasmids from white colonies using a GeneJET Plasmid Miniprep Kit (ThermoFisher Scientific, USA), based on the manufacturer's instructions, followed by *Not* I digestion. At least three clones with the correct insert size were selected for each isolate and subjected to bidirectional capillary (Sanger) DNA sequencing using primer walking at Eton Bioscience (San Diego, CA, USA).

Table 1 Analysis of predicted open reading frames in the plus-strand of the cacao mild mosaic virus and cacao yellow vein-banding virus genome sequences, using the NCBI BLASTn and BLASTp algorithms and ORF Finder tools

	ORF1		ORF2		ORF3		ORFY	
	CaMMV	CYVBV	CaMMV	CYVBV	CaMMV	CYVBV	CaMMV	CYVBV
Coordinates (nucleotides)	284 - 715	295 - 816	712 - 1104	817 - 1215	1101 - 6989	1212 - 7088	6632 - 7027	6788 - 7165
Size (nucleotides)	432	522	393	399	5589	5877	396	378
Size (amino acids)	143	173	130	132	1962	1958	131	125
% nucleotide pairwise	52 - 63%	53 - 67%	53 - 71%	54 - 63%	59 – 63%	58 – 63%	52 – 64%	52 – 63%
BLASTp % amino acid similarity^{+}	27 - 57%	23 – 43%	25 – 40%	22 – 34%	35 – 64%	34 – 69%	24 – 45%	*NV
Calculated molecular mass (kDa)	16.47	20.32	14.33	14.45	225.92	221.87	14.85	14.15
Functional conserved domain	DUF1319	DUF1319	None	None	Zn, Pep, RT, RNase H	Zn, Trim, Pep, RT, RNase H	None	None

DUF1319, domain of unknown function 1319; Zn, zinc knuckle finger; Pep, pepsin-like aspartate protease; RT, reverse transcriptase; RNase H, ribonuclease H; Trim, trimeric-dUTPase. [a] BLASTp results are based on the top 100 hits in GenBank. NV[b], No viral sequences determined

To analyze the sequences obtained using PCR and Sanger sequencing, the overlapping DNA sequences for each clone were assembled using SeqMan Pro v.12 (DNASTAR, WI, USA) and rearranged as described previously. The ORFs, functional domains, and pairwise nt sequence identities were compared with those obtained from the Illumina platform using NCBI ORF Finder, CDD, and SDT, respectively, as outlined above.

The Illumina DNA sequencing platform produced 2,111,947 and 3,664,734 reads, of which 1,084,938 and 15,355 were assembled into the cacao mild mosaic virus (CaMMV) and cacao yellow vein banding virus (CYVBV) genomes, respectively. The CaMMV genome, with 7533 bp, was assembled *de novo* from 7591 Illumina reads with a depth of coverage of 60–160. The CYVBV genome, with 7454 bp, was obtained from 2627 *de novo*-assembled reads of sequence with a depth of coverage of 20–75. The CaMMV-specific primers CaMMV_1F/1R and CaMMV_2F/2R and the CYVBV-specific primers CYVBV_1F/1R and CYVBV_2F/2R (Supplemental Table 1) were designed based on the respective Illumina-determined genome sequences. Amplification by PCR from the same purified nucleic acid preparations used for Illumina sequencing yielded amplicons of the expected size, at approximately 4 kbp for each fragment that included a 250-bp overlap at the end. The CaMMV genome sequences determined by the Illumina and Sanger methods were identical in size, 7533 bp (Fig. 1b). The Illumina CYVBV genome was 7454 bp long (Fig. 1b), whereas, the Sanger sequence was 7458 bp long, encoding an additional amino acid. The pairwise nt sequence identity for the Illumina and Sanger CaMMV sequences was 99.3%, compared to 98.8% for the CYVBV sequences. Each genome sequence was recognizable as a badnavirus by the presence of hallmark badnaviral domains and coding regions, regardless of the DNA sequencing platform. The CaMMV and CYVBV Illumina genome sequences were assigned the GenBank accession numbers KX276640 and KX276641, respectively.

Results of the cacao genome sequence searches (accession numbers FR7222157.1, CM001879.1 to CM001888.1 and KE132922.1) indicated no evidence for CaMMV or CYVBV integration.

The ORF nt coordinates and pairwise identity, aa size, molecular weight (mass), and similarity for CaMMV and CYVBV, compared with well-studied badnaviruses available in GenBank, are summarized in Table 1. Four similarly arranged ORFs were predicted for the CaMMV and CYVBV genomes, an arrangement that has been observed for most other well-studied badnaviral genomes. The CYVBV coding regions were located at nt 295–816 for ORF1, 817–1215 for ORF2, 1212–7088 for ORF3, and 6788–7165 for ORFY. The CaMMV ORF1 was located at nt 284–715; ORF2, at nt 712–1104; ORF3, at nt 1101–6989, and ORFY, at nt 6632–7027. The ORFY coding regions overlapped with the 3' end of ORF3 by 300 nt in more for both genomes.

ORF analysis indicated that CYVBV shared 52-67% nt sequence identity with other badnavirus coding regions, and BLASTp analysis showed that the encoded proteins shared 23–43% aa sequence identity for ORF1, 22–34% for ORF2, and 34–69% for ORF3, (Table 1). An analogous BLASTp analysis for the CYVBV ORFY did not reveal homology to any viral proteins for which predicted sequences are available in the GenBank database. A BLASTp comparison of CaMMV ORFs with those of other badnaviruses indicated a range of aa identity of 27–57% for ORF1, 25–40% for ORF2, 35–64% for ORF3, and 24–45% for ORFY. The CaMMV ORFs shared 52–71% overall nt sequence identity with other badnaviruses.

A search for predicted conserved protein domains in the CaMMV and CYVBV genomes implemented in the NCBI CDD [18] revealed for both viruses one domain in ORF1, an uncharacterized superfamily domain, referred to as 'domain of unknown function', or DUF1319 [5]. No conserved domains were identified in ORFs 2 and Y (Table 1). The results of CDD predictions for ORF 3 differed slightly for the Trinidad badnaviral genomes. The CaMMV ORF3 contained four domains: a zinc-finger-like RNA-binding domain associated with the coat protein and a CXCX2CX4HX4C motif [19] at nt coordinates 791–808, a pepsin-like aspartate protease domain (1199–1289), an RT-LTR domain (1419–1606), and an RNase H domain (1702 – 1830). However, the CaMMV movement protein (MP) coding region, which was expected to be the 5'-most domain in ORF3 based on other badnaviral MPs, could not be identified. Similarly, CYVBV had five domains, four of which were similar to those in CaMMV, including a zinc-finger-like RNA-binding domain at nt coordinates 810-825, a pepsin-like aspartate protease domain (1196–1284), an RT-LTR domain (1417–1603), and an RNase H domain (1699–1826). However, an additional trimeric-dUTPase-like domain with predicted involvement in hydrolysis of a dUTP-Mg complex [18] was identified in ORF3 (nt coordinates 1073–1160) of CYVBV.

Several highly conserved motifs that had been reported in well-characterized badnavirus genomes were identified in the CaMMV and CYVBV genomes. In CaMMV, there was a non-coding intergenic region with a TATA box (tacTATAAAAgga) at nt 7240-7252 and a polyA signal (AATAAA) at nt 7415-7420. Similar conserved motifs were present in the CYVBV genome: a TATA box (tatTATAAAAtaa) at nt 7376–7388 and a polyA signal (ATAAAA) at nt 7380-7385. The plant $tRNA^{Met}$ primer binding site homologs, which for CaMMV and CYVBV

were TGGTATCAGAGCTATGTT and TGGTATCAGAGCAAGGTT, respectively, were located at the nt coordinates 1–18 for both viruses.

Pairwise nt analysis of the CaMMV and CYVBV RT-RNase H region (580 bp) and the complete genome sequence, indicated 62 and 60% shared nt identity, respectively. Also, the RT-RNase H and complete genome sequences shared 59-71% and 59-62% nt sequence identity with well-characterized badnaviruses (Supplemental Tables 2 and 3). Thus, CaMMV and CYVBV are as divergent from each other as they are from all other known badnaviruses. Based on the ICTV species demarcation criterion for the genus *Badnavirus*, which is <80% nt identity, CaMMV and CYVBV should be regarded as members of new species (Supplemental Table 2).

A phylogenetic tree constructed based on 84 RT-RNase H and complete badnavirus genome sequences placed CaMMV and CYVBV in different groups. Also, the RT-RNase H and complete genome trees were incongruent, in that CYVBV grouped with the same closest relatives in both trees, while CaMMV grouped with different closest relatives in each tree (Fig. 2a, b). Further, the RT-RNase H tree resolved a polytomy, but with poor bootstrap support (Fig. 2a). Although the complete genome sequence is not used for taxonomic considerations, it had three well-supported clades with 79-100% bootstrap values, referred to as clades 1-3 (Fig. 2b.). Based on the complete-genome tree, CaMMV grouped in clade 2 with nine other badnaviruses, including the cacao-infecting CSSV from West Africa. In contrast, CYVBV grouped in clade 3 with nine other badnaviruses (Fig. 2b) from diverse plant species, none of which are cacao. Clade 1 contained non-cacao-infecting badnaviruses that have thus far been associated with plant families originating in Africa, Asia and the Pacific, while clade 2 badnaviruses have been reported to infect plants endemic to Asia, and Central or South America, and the viruses in clade 3 are associated with plants originating in Africa and Europe [11, 21].

Here, the genome sequences of the previously elusive CaMMV and CYVBV (formerly strain A and B, respectively) were determined using Illumina HiSeq for discovery and confirmed by PCR amplification and primer walking for each recovered, cloned viral genome. Because the cacao plants in which these viruses were found showed symptoms similar to those observed for strains A and B, it is likely that CaMMV and CYVBV were the suspect viral pathogens causing foliar "red mottling" and "vein-clearing" symptoms, respectively, in cacao in Trinidad during the 1940s [17]. Until this report, only one badnavirus, CSSV, the causal agent of swollen shoot disease of cacao, had been identified from *T. cacao* plants, and thus far, it has been found only in West Africa.

Consistent with other badnaviral genomes, CaMMV and CYVBV contain the 'badnaviral core' ORFs 1-3 and a fourth, referred to as ORFY. The genome of CSSV, the only cacao-infecting badnavirus previously described until this report, also contains an ORFY that overlaps with the 3'end of ORF3, making it analogous to ORFY in CaMMV and CYVBV. However, the CaMMV and CYVBV genomes contained no discernable ORFX like that found within ORF3 of six out of the seven CSSV genomes, for which no function has been ascribed. Overall, the genome structure and ORF arrangement for CaMMV and CYVBV are similar to those of other badnaviruses, some of which also contain ORFY [6, 12].

The intergenic region (IR) of CaMMV and CYVBV contains several functionally conserved motifs of interest. Both viral genomes contain a $tRNA^{Met}$ binding site that occurs in the plant host and serves to initiate viral RNA transcription [19]. Also in the IR are the polyadenylation signal and TATA box motifs that are retained within terminally redundant full-length, badnavirus transcripts [23]. The presence of these hallmark features further support the identification of CaMMV and CYVBV as badnaviruses. Also, the CaMMV and CYVBV ORF arrangement and respective ORF sizes are consistent with those of other pararetroviruses (Table 1) [16].

Because CaMMV and CYVBV share 60% nt sequence identity, they are as divergent from each other as they are from other badnaviruses, including CSSV from West Africa, the only other badnavirus known to infect cacao. Because the CaMMV and CYVBV RT-RNase H loci share only 57-71% nt sequence identity with those of other badnaviruses and the badnavirus species cutoff is <80%, these viruses are considered members of distinct species [16].

The re-emergence of the disease caused by these viruses in Trinidad is not likely attributable to activation of existing endogenous badnaviruses, since we found no evidence for integration events in the cacao genome sequence. Even so, direct analysis of genomic DNA from CaMMV- and CYVBV-infected cacao plants is needed to provide more-robust proof of the lack of integration, for example, using RCA and restriction digestion [15] or Southern hybridization.

Foliar symptoms associated with CaMMV- and CYVBV-infected cacao trees in Trinidad are like those observed in CSSV-infected cacao in West Africa, but isolates causing shoot swelling occur uniquely in West Africa, strongly suggesting that at least certain mechanisms of pathogenicity differ between badnaviruses endemic to Africa and those extant in the New World. Whether the ancestors of CaMMV and CYVBV and/or primary hosts of the parental viruses are endemic to the Eastern Caribbean region or elsewhere is not known. If it was an exotic

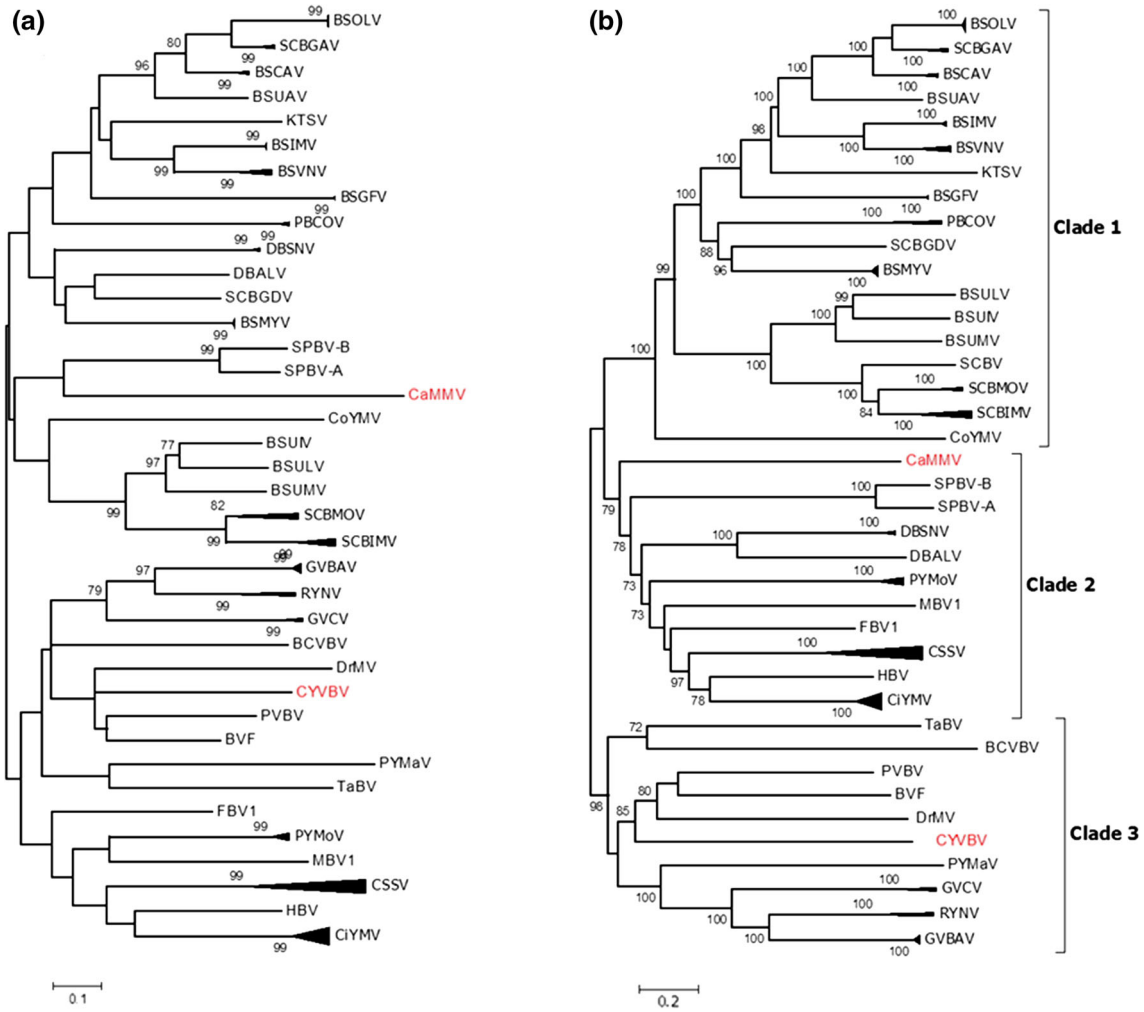

Fig. 2 Maximum-likelihood (ML) tree reconstructed for 35 badnaviruses, including cacao mild mosaic virus and cacao yellow vein-banding virus, based on the 580-bp RT-RNase H region (a) and the complete badnaviral genome sequence (b). The ML analysis was carried out using MEGA 6 software for 1000 iterations. The horizontal branch lengths are proportional to the genetic distance. Numbers shown at branch points indicate bootstrap values >70%. Each node was collapsed into a single representative taxon when more than one sequence was available for a particular species. Trees were rooted at the midpoint. PYMaV, pagoda yellow mosaic associated virus (NC_024301); TaBV, taro bacilliform virus (NC_004450); SPBV-A-sweet potato badnavirus A (NC_015655); SPBV-B, sweet potato badnavirus B (NC_012728); MBV1, mulberry badnavirus 1 (NC_026020); DrMV, dracaena mottle virus (NC_008034); CoYMV, commelina yellow mottle virus (NC_001343); BCVBV, Bougainvillea spectabilis chlorotic vein-banding virus (NC_011592); SCBIMV, sugarcane bacilliform IM virus (NC003031, JN377533, JN377536, JN377537); SCBMOV, sugarcane bacilliform Mor virus (NC_008017, JN377534); SCBV, sugarcane bacilliform virus (JN377535); SCBGAV, sugarcane bacilliform Guadeloupe A virus (FJ824813, FJ824814); SCBGDV, sugarcane bacilliform Guadeloupe D virus (NC_013455); BSUMV, banana streak UM virus (NC_015505); BSUIV, banana streak UI virus (NC_015503); BSULV, banana streak UL virus (NC_015504); CSSV, cacao swollen shoot virus (NC_001574, AJ609019, AJ608931, AJ609020, AJ534983, JN606110, AJ781003); PYMoV, piper yellow mottle virus (NC_022365, KJ873041, KJ873042, KJ873043); GVBAV, gooseberry vein banding associated virus (NC_018105, HQ852248, HQ852249, HQ852250, HQ852251); RYNV, rubus yellow net virus (NC_026238, KF241951); HBV, hibiscus bacilliform virus (NC_023485); FBV1, fig badnavirus 1 (NC_017830); CiYMV, citrus yellow mosaic virus (NC_003382, FJ617224, JN006805, JN006805, EU708316, EU708317, JN006805, EU489744, EU489745); PVBV, pelargonium vein banding virus (NC_013262); BVF, blackberry virus F (NC_029303); PBCOV, pineapple bacilliform comosus virus (NC_014648, GQ398110); GVCV, grapevine vein clearing virus (NC_015784.2, KJ725346); BSGFV, banana streak GF virus (NC_007002, KJ013507); BSMYV, banana streak MY virus (NC_006955, KF724854, KF724855, KF724856, EU140339, KJ013509); DBSNV, dioscorea bacilliform SN virus (NC_009010, DQ822074); DBALV, dioscorea bacilliform AL virus (KF829952); KTSV, kalanchoe top spotting virus (NC_004540); BSUAV, banana streak UA virus (NC_015502); BSOLV, banana streak OL virus (NC_003381, KJ013506, JQ409540, JQ409539, DQ859899, DQ451009); BSCAV, banana streak CA virus (NC_015506, KJ013511); BSIMV, banana streak IM virus (NC_015507, KJ013508, HQ659760); BSVNV, banana streak Vietnam acuminata virus (KJ013510, NC_007003); CaMMV, cacao mild mosaic virus (KX276640); CYVBV, cacao yellow vein-banding virus (KX276641)

introduction, one possible source could have been virus-infected cacao from its center of diversity in the Amazon Basin or the area where it has been domesticated in the region [21]. Even so, cacao-infecting badnaviruses have not been previously reported in the American Tropics. Alternatively, a host shift to cacao by CaMMV- and CYVBV-like ancestors either endemic to or recently introduced into Trinidad could have resulted in the first outbreaks recorded over 80 years ago. Another possibility is that CaMMV- and CYVBV-like viruses could have been introduced into Trinidad with exotic plants; indeed, many badnaviruses have been associated with plant hosts that were once considered exotic, but now are naturalized in the Eastern Caribbean region [11].

Compliance with ethical standards

Funding We gratefully acknowledge funding for this project from the USDA-ARS through the Specific Cooperative Agreement #6038-21000-023-07 titled: Development and Optimization of Molecular Diagnostics Method for Qualitative and Quantitative Detection of Cacao Swollen Shoot Virus with MARS Inc through Trust Agreement #58-6631-6-123, titled: Genetic Improvement of Theobroma cacao, and from the World Cocoa Foundation Borlaug Fellowship Program, USDA-Foreign Agricultural Services.

Conflict of interest The authors declare that they have no potential conflict of interest.
This manuscript does not contain studies with human participants or animals.

References

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410
2. Argout X, Salse J, Aury J, Guiltinan MJ, Droc G et al (2011) The genome of *Theobroma cacao*. Nat Genet 43:101–108. doi:10.1038/ng.736
3. Baker RED, Dale WT (1947) Notes on a virus disease of cacao. Ann Appl Biol 34:60–65. doi:10.1111/j.1744-7348.1947.tb06343.x
4. Baker RED, Dale WT (1947) Virus diseases of cacao in Trinidad II. Trop Agric 24:127–130
5. Bateman A, Coggill P, Finn RD (2010) DUFs: Families in search of function. Acta Crystallogr Sect F Struct Biol Cryst Commun 66:1148–1152. doi:10.1107/S1744309110001685
6. Borah BK, Johnson AMA, Sai Gopal DVR, Dasgupta I (2009) Sequencing and computational analysis of complete genome sequences of Citrus yellow mosaic badna virus from acid lime and pummelo. Virus Genes 39:137–140. doi:10.1007/s11262-009-0367-9
7. Conesa A, Gotz S (2008) Blast2GO: A comprehensive suite for functional analysis in plant genomics. Int J Plant Genomics. doi:10.1155/2008/619832
8. Dean FB, Nelson JR, Giesler TL, Lasken RS (2001) Rapid amplification of plasmid and phage DNA using Phi29 DNA polymerase and multiply primed rolling circle amplification. Genome Res 11:1095–1099. doi:10.1101/gr.180501.4
9. Doyle JJ, Doyle JL (1990) Isolation of plant DNA from fresh tissue. Focus (Madison) 12:13–15
10. Gray A (2001) The world cocoa market outlook. LMC International Ltd., London
11. Hancock JF, Miller AJ (2014) Crop plants: Evolution. In: eLS. Wiley, Chichester. doi:10.1002/9780470015902.a0003360.pub2
12. Hany U, Adams IP, Glover R, Bhat AI, Boonham N (2014) The complete genome sequence of Piper yellow mottle virus (PYMoV). Arch Virol 159:385–388. doi:10.1007/s00705-013-1824-2
13. Jacquot E, Hagen LS, Jacquemond M, Yot P (1996) The open reading frame 2 product of Cacao swollen shoot badnavirus is a nucleic acid-binding protein. Virology 225:191–195
14. Jacquot E, Hagen LS, Michler P, Rohfritsch O, Stussi-Garaud C, Keller M, Jacquemond M, Yot P (1999) In situ localization of Cacao swollen shoot virus in agroinfected Theobroma cacao. Arch Virol 144:259–271
15. James AP, Geijskes RJ, Dale JL, Harding RM (2011) Development of a novel rolling-circle amplification technique to detect Banana streak virus that also discriminates between integrated and episomal virus sequences. Plant Dis 95:57–62. doi:10.1094/PDIS-07-10-0519
16. King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ (2012) Badnavirus. In: Virus taxonomy. Ninth report of the international committee on taxonomy of viruses. Elsevier Academic Press, London
17. Kirkpatrick TW (1950) Insect transmission of cacao virus disease in Trinidad. Bull Entomol Res 41:99. doi:10.1017/S0007485300027504
18. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Lu F, Marchler GH, Mullokandov M, Omelchenko MV, Robertson CL, Song JS, Thanki N, Yamashita RA, Zhang D, Zhang N, Zheng C, Bryant SH (2011) CDD: a conserved domain database for the functional annotation of proteins. Nucleic Acids Res 39:225–229. doi:10.1093/nar/gkq1189
19. Medberry SL, Lockhart BEL, 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 18:5505–5513
20. Motamayor JC, Mockaitis K, Schmutz J, Haiminen N, Iii DL, Cornejo O, Findley SD, Zheng P, Utro F, Royaert S, Saski C, Jenkins J, Podicheti R, Zhao M, Scheffler BE, Stack JC, Feltus FA, Mustiga GM, Amores F, Phillips W, Marelli JP, May GD, Shapiro H, Ma J, Ma J, Bustamante CD, Schnell RJ, Schnell RJ, Main D, Gilbert L, Parida L, Kuhn DN (2013) The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. Genome Biol 14:r53. doi:10.1186/gb-2013-14-6-r53
21. Motamayor JC, Risterucci AM, Lopez PA, Ortiz CF, Moreno A, Lanaud C (2002) Cacao domestication I: the origin of the cacao cultivated by the Mayas. Heredity (Edinb) 89:380–386. doi:10.1038/sj.hdy.6800156
22. Muhire BM, Varsani A, Martin DP (2014) SDT: a virus classification tool based on pairwise sequence alignment and identity calculation. PLoS One 9:e108277. doi:10.1371/journal.pone.0108277
23. Muller E, Sackey S (2005) Molecular variability analysis of five new complete cacao swollen shoot virus genomic sequences. Arch Virol 150:53–66. doi:10.1007/s00705-004-0394-8
24. Posnette AF (1944) Viruses of cacao in Trinidad. Trop Agric 21:105–106
25. Rector A, Tachezy R, Van Ranst M (2004) A sequence-independent strategy for detection and cloning of circular DNA virus genomes by using multiply primed rolling-circle amplification. J Virol 78:4993–4998. doi:10.1128/JVI.78.10.4993

26. Sreenivasan TN (2009) The enigma of the ICS 76 plants at Reading. UK, Reading, UK
27. Stevens H, Rector A, Van Ranst M (2010) Multiply primed rolling-circle amplification method for the amplification of circular DNA viruses. Cold Spring Harb Protoc 2010:pdb.prot5415. doi:10.1101/pdb.prot5415
28. Swarbrick JT (1961) Cacao virus in Trinidad. Trop Agric 38:245–249
29. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 30:2725–2729
30. Thorold CA (1975) Diseases of cacao. Clarendon Press Oxford 77–78
31. Villesen P (2007) FaBox: An online toolbox for FASTA sequences. Mol Ecol Notes 7:965–968. doi:10.1111/j.1471-8286.2007.01821.x
32. Yang IC, Hafner GJ, Revill PA, Dale JL, Harding RM (2003) Sequence diversity of South Pacific isolates of Taro bacilliform virus and the development of a PCR-based diagnostic test. Arch Virol 148:1957–1968. doi:10.1007/s00705-003-0163-0