Bioinformatic, genetic and molecular analysis of several badnavirus sequences integrated in the genomes of diverse cocoa (Theobroma cacao L.) germplasm

Saudi Journal of Biological Sciences 30 (2023) 103648

HOSTED BY

Contents lists available at ScienceDirect

Saudi Journal of Biological Sciences

journal homepage: www.sciencedirect.com

Original article

Bioinformatic, genetic and molecular analysis of several badnavirus sequences integrated in the genomes of diverse cocoa (*Theobroma cacao* L.) germplasm

Ihsan Ullah, Jim M. Dunwell *

School of Agriculture, Policy and Development, University of Reading, Reading RG6 6EU, UK

ARTICLE INFO

Article history:
Received 9 February 2023
Revised 14 March 2023
Accepted 31 March 2023
Available online 11 April 2023

Keywords:
Badnavirus
Cocoa
EVEs

ABSTRACT

Endogenous viral elements (EVEs) are integrations of whole or partial viral genomes into the host genome, where they act as host alleles. They exist in a wide range of plant species including *Theobroma cacao*, the source of chocolate. Because of the international transfer of cacao germplasm, it is important to discriminate between the presence of these inserts and any episomal viruses that may be present in the material. This study was designed to survey a wide range of cacao germplasm, to assess the number, length, orientation, and precise location of the inserts and to identify any effect on the transcription of the gene into which they are inserted. Using a combination of bioinformatic, genetic and molecular approaches, we cloned and sequenced a series of different inserts, including one full-length virus sequence. We also identified, for the first time, an inhibitory effect of the insert on the expression of host genes. Such information is of practical importance in determining the regulation of germplasm transfer and of fundamental relevance to aiding an understanding of the role that such inserts may have on the performance of the host plant.

1. Introduction

Endogenous viral elements (EVEs) are integrations of whole or partial viral genomes into the genome of their hosts and are then inherited as host alleles (Katzourakis and Gifford, 2010). Integration of retroviruses (RNA-retrotranscribing viruses) into the host genomes is an obligatory step for their replication, hence endogenization of remnants of ancient retroviruses in animal and human genomes is prevalent (Johnson, 2019; Li et al., 2022). Plants lack true retroviruses; however, double-stranded (ds)DNA reverse-transcribing plant viruses (Pararetroviruses) from the *Caulimoviridae* family contain many common retroviral features and are found endogenized in several plant species. Unlike retroviruses, these dsDNA do not encode an integrase and replicate without integration in the host nuclear genome through an obligatory RNA intermediate, followed by reverse transcription to complete the life cycle (Pfeiffer and Hohn, 1983). Several possible mechanisms have been suggested for integration of such caulimovirid sequences in host genomes; these include use of host's integrase, non-homologous to known integrase enzymes (Richert-Pöggeler et al., 1997); hijacking the integrases encoded by other retrotransposons (Richardson et al., 2015); and microhomology-mediated recombination or nonhomologous end-joining (Brown, 2003).

The EVEs exist in a wide range of plant species (Boutanaev and Nemchinov, 2021; Schmidt et al., 2021; Bhat et al., 2022), mostly in the form of partial repetitive fragments dispersed in the host genome without any deleterious effect on their hosts. However, endogenization of functional full-length viral genomes can trigger systemic infection of the host plant. Examples include *Banana streak virus* (*Badnavirus)* in banana (Ndowora et al., 1999; Gayral et al., 2008), *Petunia vein clearing virus* (*Petuvirus)* in petunia (Richert-Pöggeler et al 1997), and *Tobacco vein clearing virus* (*Solendovirus*) in tobacco (Lockhart et al., 2000; Gregor et al., 2004). Regardless of their size, EVEs inserted into the coding sequence of the host genes are likely to impact the function of the host genes (Serfraz et al., 2021). Moreover, a recent report in tomato found that the EVEs-derived small (s)RNAs can alter patterns of gene expression of the host plant (Lopez-Gomollon et al., 2022).

* Corresponding author.
E-mail addresses: i.ullah@reading.ac.uk (I. Ullah), j.m.dunwell@reading.ac.uk (J.M. Dunwell).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.sjbs.2023.103648
1319-562X/

Cocoa (*Theobroma cacao* L.) is an economically important plant species, mainly cultivated for beans in humid tropical areas in West Africa, Central and South America, and South-East Asia. West Africa is the largest cocoa producing region, accounting for 77% of world cocoa production (https://www.icco.org/wp-content/uploads/Production_QBCS-XLVIII-No.-2.pdf). Great diversity exists in cacao accessions, which are categorized in two broad genetic groups "Criollo" and "Forastero" on the basis of morphological and geographical origins. A third group "Trinitario" consists of germplasm derived from hybridization of "Criollo" X "Forastero" accessions. Recently, a more precise classification of cacao based on microsatellite markers divided cacao into ten genetic groups Amelonado, Contamana, Criollo, Curaray, Guiana, Iquitos, Marañon, Nacional, Nanay, and Purús (Motamayor et al., 2008).

Badnaviruses are known to infect cacao, with ten such viruses recognized by the International Committee on Taxonomy of Viruses (ICTV). Specifically, seven species associated with Cacao swollen shoot virus (CSSV) disease are prevalent in West Africa and cause a significant loss of productivity. Another species, Cacao bacilliform Sri Lanka virus (CBSLV) has been reported in Sri Lanka (Muller et al., 2018). In addition, two new badnavirus species designated as cacao mild mosaic virus (CaMMV) and cacao yellow vein-banding virus (CYVBV), which cause mild symptoms, have been identified in cacao trees grown in the western hemisphere (Chingandu et al., 2017; Puig, 2021). These two species have also recently been reported in asymptomatic trees in cacao germplasm in the United Kingdom (Ullah et al., 2021). CaMMV has also recently been detected in South-East Asia (Kandito et al., 2022).

In addition to the cacao badnavirus species, EVEs of other badnaviruses have been found in cacao genomes. Initially, badnavirus-like sequences were identified in cacao trees in West Africa in 2018 and were designated as CSS Ghana S virus isolates (Muller et al., 2018). Later, we demonstrated that five of these isolates are in fact integrated badnaviral sequences prevalent in the host genomes and varied in type, with each type predominating in a specific cacao genetic group. One of the five inserts was further studied to determine the length of the insert, chromosomal location, zygosity level in specific cacao accessions, and inheritance pattern (Muller et al., 2021). These sequences were designated as endogenous *T. cacao* bacilliform virus 1 (eTcBV1). In a significant extension to the previous work, here we report a comprehensive analysis of three additional eTcBV1 sequences. Bioinformatic, molecular, and genetic approaches were employed for characterization of these inserts. We also report for the first time an eTcBV1 that contains a complete badnavirus genome. In addition, the impact of the insertions on expression of their host genes is demonstrated.

2. Materials and methods

2.1. Analysis of genomic and transcriptome datasets

The raw read FastQ format files of the third party genomic and transcriptome datasets comprising PRJNA734904 (Osorio-Guarín et al., 2020) and PRJNA558793 (Hämälä et al., 2021) were downloaded from the Sequence Read Archive (SRA, https://trace.ncbi.nlm.nih.gov/Traces/sra/?run) and the European Nucleotide Archive (ENA, https://www.ebi.ac.uk/ena/browser/home). The 229 genomic datasets downloaded from BioProject PRJNA734904 were searched for eTcBV1s using a refence database consisting of five badnavirus S sequences (Supplementary Table S1), whereas 31 RNA-seq datasets downloaded from BioProject PRJNA558793 were used to determine co-expression of cacao *TcRGA3, TcPAP26* and *TcAAR2* genes. The Bowtie2 v 2.3.4.1 aligner (Langmead and Salzberg, 2012) was employed to map the read to the reference databases. The mapped reads were compressed, sorted and indexed by Samtools v 1.10 (Li et al., 2009). The alignment data were visualised in the Integrative Genomics Viewer (IGV) v.2.4.13 (Robinson et al., 2017). *De novo* assemblies of selected cacao accessions were downloaded from the Penn State University website (https://bigdata.bx.psu.edu/Cacao_NSF_data/) and searched for positive contigs that contained eTcBV1 appended with cacao sequence. Blast searches were conducted using the positive contig sequence as a query against the cacao genome in order to identify the precise insertion site. For confirmation, a Sanger sequence contig of eTcBV1-II, SeqMan Ultra 17.3 with option of "combined reference-guided *de novo* assembly" was employed to construct an assembly from Illumina dataset SRR9938225 using the 10,170 nt contig formed from PA107 (eTcBV1-II) as reference assembly.

2.2. Plant material

A fully expanded young leaf was used for DNA or RNA extraction from cacao accessions held at International Cocoa Quarantine Centre, Reading, UK (ICQC-R). Controlled self-pollinations were conducted at ICQC-R using the accessions that were found hemizygous for eTcBV1-I, II or III locus. Following five to six months of pod development, the resultant seeds were collected and germinated.

2.3. Nucleic acid extraction

One hundred milligrams of a sample were ground with liquid nitrogen in a microcentrifuge tube in the presence of tungsten carbide beads using a TissueLyser II (Qiagen, UK). Total genomic DNA was extracted using the DNeasy 96 plant kit (Qiagen, UK), whereas total RNA was isolated by a modified CTAB method (see supplementary file for detailed protocol) followed by purification by the RNeasy Plant Mini Kit (Qiagen, UK) according to the manufacturer's instructions. The quality and quantity of isolated nucleic acids were determined using a NanoDrop 2000 Spectrophotometer (Thermo Fisher Scientific, UK). Aliquots of extracted nucleic acids were also run on a 1–1.5% agarose gel for quality analysis.

2.4. PCR amplification

Multiple primer sets were designed with online tool Primer3-Plus (https://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi) based on the information obtained from *de novo* assemblies about insertion site of viral sequences in the cacao genome. The details on insertion sites and location of the primers are provided in Figs. 1 and 2. The Phusion green hot start II high-fidelity PCR master mix was used to amplify the viral insert along with the bordering host sequence. The PCR reaction, which contained 25 μL of PCR Mix, 4 μL each of 5 μM forward and reverse primers, 5 μL of 10 ng/μl DNA template and 12 μL of PCR water, was performed in a thermal cycler (Veriti, Applied Biosystems, UK) programmed to one cycle of 98 °C for 1 min, followed by 35 cycles of 98 °C for 10 s, 63 °C for 10 s and 72 °C for 2 min. Final extension was performed at 72 °C for 10 min. The amplicon was cloned with Zero Blunt™ TOPO™ PCR Cloning Kit for Sequencing. The cloned PCR products were sequenced by Sanger technology (Source Bioscience, UK). The information obtained from sequencing of the amplified viral inserts was then utilized to design multiplex PCR assays for screening of ICQC-R germplasm. The multiplex PCR reaction consisting of 10 μL of Platinum Hot Start PCR Master Mix, 1.5 μL each of 5 μM four primers, 2 μL of 10 ng/μL DNA template and 2 μL of PCR water was performed in the thermal cycler programmed to one cycle of initial denaturing at 94 °C for 2 min, followed by 30 cycles of denaturing at 94 °C for 15 s, primer annealing at 58/60/63 °C for 15 s and extension at 68 °C for 20 s.

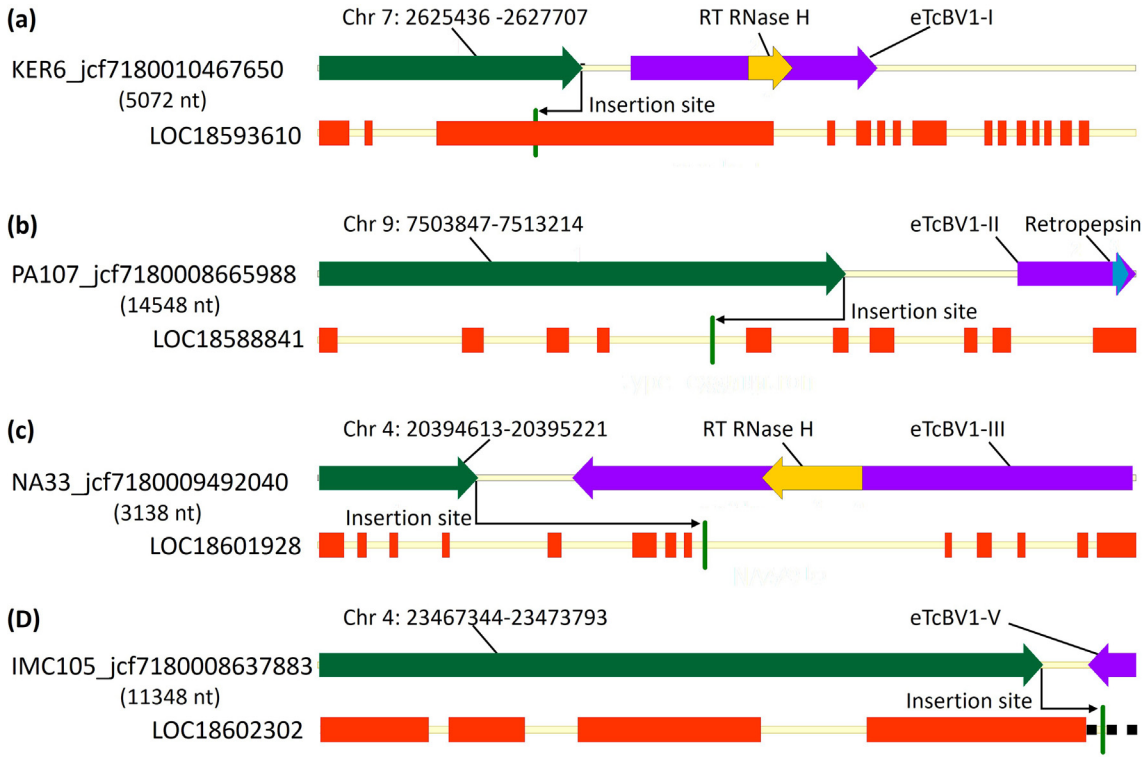

Fig. 1. Insertion sites of endogenous *T. cacao* bacilliform virus1 (eTcBV1) found in cacao genomes. Preliminary assemblies of cacao genomes generated from the data downloaded from BioProject PRJNA558793 were searched for contigs that contained eTcBV1 with bordering cacao genomic region. Purple and green arrows represent eTcBV1 and bordering genomic region of cacao (B97 *T. cacao* reference genome), respectively. Yellow and blue arrows denote RT RNase H and Retropepsin conserved domains, respectively. Each contig structure is followed by the structure of the gene having an eTcBV1. Red bars and yellow horizontal lines represent exons and introns, respectively. The insertion site in a gene is denoted by a vertical green line. (a) Insertion site of eTcBV1-I in disease resistance protein RGA3 (LOC18593610). (b) Insertion site of eTcBV1-II in bifunctional purple acid phosphatase 26 (LOC18588841). (c) Insertion site of eTcBV1-III in protein AAR2 homolog (LOC18601928). (d) Insertion site of eTcBV1-V in pre-mRNA-splicing factor CWC25 homolog (LOC18602302). Black dotted horizontal line indicates intergenic region between genes LOC18602302-LOC18602303.

The PCR products were resolved on a 1.5 % agarose gel and stained with ethidium bromide. Information about the primers used in this study including primer sequences, specific targets and annealing temperatures for PCR is provided in Table 1.

2.5. Sequence analysis and phylogeny

The sequences of 59 reference genomes of virus species classified in the genus Badnavirus in ICTV, and 87 selected genomes representing 10 cacao-infecting badnaviruses were downloaded from NCBI (https://www.ncbi.nlm.nih.gov/). The putative open reading frames (ORFs), predicted amino acid sequence, and molecular weight of the encoded proteins of the downloaded sequences and endogenous sequences found in this study were identified using the CLC sequence viewer (Qiagen, UK). Protein domain features were predicted using InterPro database (https://www.ebi.ac.uk/interpro/) (Blum et al., 2021). Molecular Evolutionary Genetics Analysis (MEGA) version 11 (Tamura et al., 2021) was used for calculation of Relative Synonymous Codon Usage (RSCU) and evolutionary analysis Maximum Likelihood tree was constructed for 128 eTcBV1 positive cacao accessions based on distance matrix using 151 SNP markers. The RSCU of cacao was downloaded from codon statistics database (https://codonstatsdb.unr.edu/).

2.6. Expression analysis

Isolated total RNA (1.0 µg) was converted into cDNA using SuperScript™ IV VILO™ with ezDNase™ Enzyme Master Mix (Thermo Fisher Scientific, UK). Primers were designed (Table 1) from cacao genes that contain insertions of eTcBV1-III (*TcAAR2;* GenBank accession LOC18601928) and eTcBV1-II (*TcPAP26*; GenBank accession LOC18588841) and the reference gene ACP1 (LOC18599903) using Primer Blast tool (https://www.ncbi.nlm.nih.gov/tools/primer-blast). StepOnePlus™ Real-Time PCR system was used for real time RT-PCR. PowerUp™ SYBR® Green master mix (Applied Biosystems, UK) was used in qRT-PCR. The $2^{-\Delta\Delta Ct}$ method was used to determine the relative changes in gene expression. The acyl carrier protein *TcACP1* gene was utilised for normalization of expression.

3. Results

3.1. Identification of eTcBV1 in genomic sequence datasets

We have previously reported the prevalence of five types of eTcBV1s in 243 whole genome sequencing (WGS) datasets from five different studies, and the association of each type of eTcBV1s with a specific cacao genetic group (Muller et al., 2021). Recent availability of genomic sequence data of cacao germplasm held at Agrosavia, Colombia (Osorio-Guarín et al., 2020) provides an excellent opportunity to extend the search for eTcBV1 in this diverse collection. The data in this study were generated by sequencing reduced representation libraries representing 229 cacao accessions. We identified eTcBV1s in datasets from 64 accessions, with a single eTcBV1 detected in 48 accessions, and multiple eTcBV1s in 16 accessions (**Supplementary Fig. S1**). eTcBV1-VI was found to be the most prevalent, detected alone in 22 datasets, followed by eTcBV1-III, eTcBV1-II and eTcBV1-V that were individually found in 14, 10 and 2 datasets, respectively. We did not find eTcBV1-I alone in the datasets studied, a finding that is not surpris-

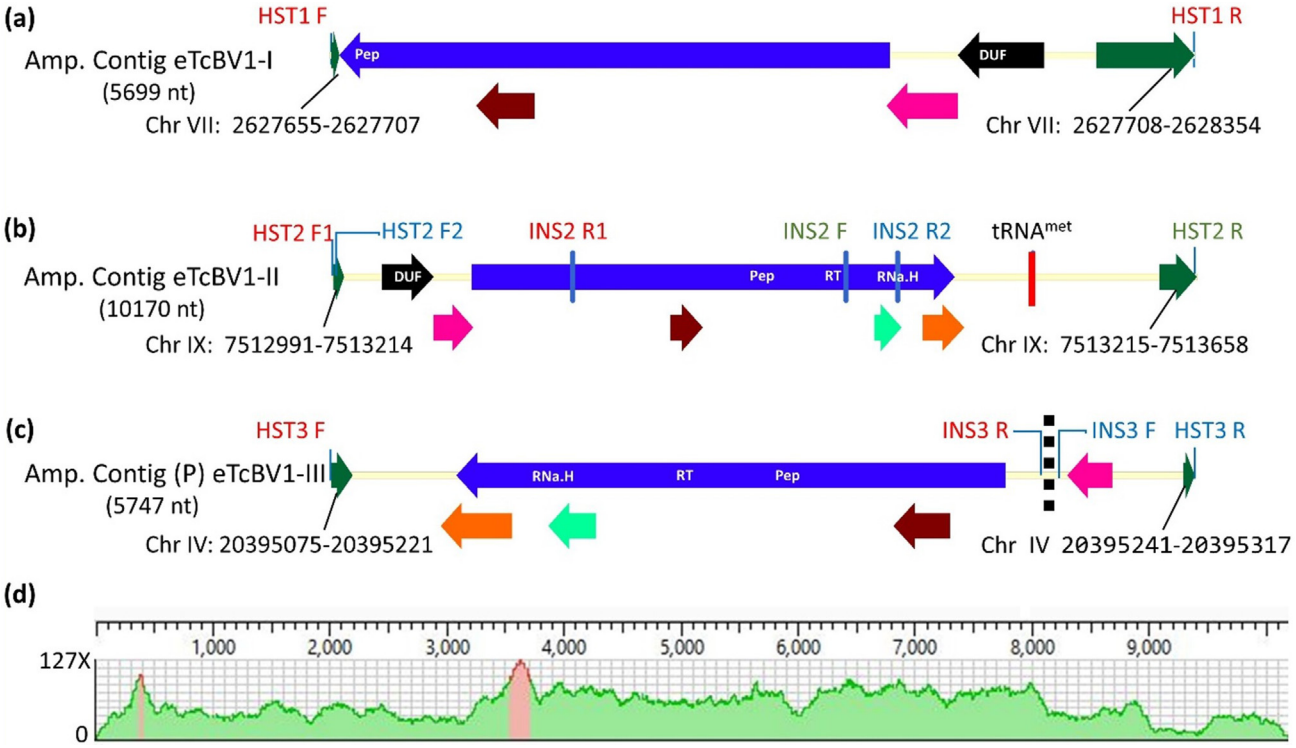

Fig. 2. Amplification of endogenous *T. cacao* bacilliform virus1 (eTcBV1) from cacao clones. The eTcBV1 types I, II and III along with the bordering host sequence were amplified from cacao accessions. Vertical blue lines indicate location of primers used in amplification. For each type, same font colour for primer names represents a pair of primers. Green arrows on both sides represent genomic region of cacao (B97 *T. cacao* reference genome) bordering an eTcBV1. Black, pink, blue, dark brown, light green and light brown arrows denote opening reading frame, 1, 2, 3, 4, 5 and Y, respectively. Predicted conserved domains comprising Domain of Unknown Function (DUF1319), pepsin-like aspartate protease (Pep) reverse transcriptase (RT) and ribonuclease H (RNa.H) are shown on the ORFs (a) A 5699 nt long contig having eTcBV1-I constructed from accession KER6. (b) A 10,170 nt long contig having eTcBV1-II constructed from accession PA107. Red vertical line indicates location of $tRNA^{Met}$ binding site. (c) A 5747 nt contig having eTcBV1-III constructed from accession NA33. Dotted vertical black line indicates a gap (d) Coverage depth of the contig generated from third party Illumina data (SRR9938225) for validation of the eTcBV1 types II contig assembled from Sanger sequencing of overlapping amplicons.

ing, as the selected germplasm lacks accessions from the Guiana genetic group where this type has been predominantly found. Among the multiple eTcBV1s found in 16 datasets, six datasets had a combination of eTcBV1-I and eTcBV1-II, and four contained both eTcBV1-III and eTcBV1-VI. Two datasets had eTcBV1-II and eTcBV1-V, and accession SUI69 hosted three types of eTcBV1. We previously found 52% prevalence of eTcBV1s in WGS genomes. Here we found 27% of datasets that contained eTcBV1s, a lower frequency that is to be expected considering the nature of the data, with low coverage of the genome and the type of germplasm. The correlation between the genetic grouping of the cacao accessions and type of eTcBV1 mapped was also consistent with previous findings (**Supplementary Table S2**). Marañon accessions contained eTcBV1-II, whereas Nanay accessions had the eTcBV1-III sequence. Similarly, the eTcBV1-V sequence was prevalent in accessions referred to as Iquitos reference. The eTcBV1-VI sequence was prevalent in the accessions with admixed ancestry. There are 19 accessions common to this study and the three BioProjects previously analysed. Findings for 18 of these accessions are in agreement among the studies (**Supplementary Table S3**). The exception was accession EET59, which was found positive for eTcBV1-VI previously, though not in the present study. This finding, again, may be attributed to the very small amount of sequence data available (158 Mb).

3.2. Identification and characterizing of eTcBV1s in cacao accessions

Based on evidence from the bioinformatic analyses of genomic sequence datasets in multiple studies, and availability of the germplasm in the Reading collection, the *de novo* assemblies of cacao accessions PA107 [PER], NA33, GU123/V and IMC67 were searched for the contig that mapped to the targeted eTcBV1 together with the *T. cacao* sequence. We found positive contigs in three datasets. The eTcBV1s fragments were amplified, sequenced and analysed. Results are described in sequence, from identification of an integrant that comprises a complete badnavirus genome, to those that were found as shorter sequences.

3.2.1. eTcBV1-II

A 14548 bp contig found in the PA107 [PER] assembly contains the eTcBV1-II reference sequence from position 10300–12400 nt. The contig region from position 1–9378 nt is 99% identical with the Chromosome IX region 2625436–2627707 nt (GenBank accession NC_030858), with the putative insertion site being in intron 4 of the *TcPAP26* gene, which encodes the bifunctional purple acid phosphatase 26 (GenBank accession LOC18588841) (Fig. 1**b**).

Data obtained from the *de novo* assembly searches were utilized to design multiple primer pairs from the bordering host sequence and known insert sequence, for amplification of full length eTcBV1s and verification of their insertion sites in the cacao genome (Fig. 1, Table 1). No successful amplification of full-length sequence was found from the accession PA107 [PER] (Reading accession RUQ 1728), with the host border primers. Therefore, the eTcBV1-II locus was amplified in three overlapping fragments with nested PCR.

Two fragments were amplified targeting 3839 and 4138 bp regions of the left border host and insert sequence, and the right border host and insert sequence, respectively. Both amplicons were cloned and sequenced. A third 6486 bp fragment was amplified with the left primer designed from the host sequence region

Table 1
PCR primers used in this study.

Specificity	Primer name	Primer location	Primer sequence	Ta (°C)	Amplicon size (bp)
Insert along with flanking host (cacao) sequence					
eTcBV1-I	HST1 F	Host left border	TGGAAGGTGCTAGAAAAAGATGA	63	5699
	HST1 R	Host right border	CAAGACTGGTGCAATGTTCG		
eTcBV1-II	HST2 F1	Host left border	ACCCAAATCACCTGCAACAT	63	3839
	INS2 R1	Insert	TTTCCGGGATATGCAATGTT		
	INS2 F	Insert	AGCATTCCAACGAAAGATGG	63	4138
	HST R	Host right border	AGGCAAGTTCCTGGTTTCAA		
	HST2 F2	Host left border	CCACAAAAGATGAGAGAGAGACC	63	6486
	INS2 R2	Insert	ATGGTTGCTTTTGGAACAGG		
eTcBV1-III	HST3 F	Host left border	TTCTGTTTGGCTGTACAGAGG	63	4651
	INS R	Insert	TGGCTCCAGTGGAAGTTAGC		
	INS3 F	Insert	ATTTGCGCAGTCCGTATTTC	63	1098
	HST3 R	Host right border	GAAGTCAAAGGGGGAAAAGG		
Genotyping					
eTcBV1-I	1HST F	Host left border	TGGAAGGTGCTAGAAAAAGATGA	60	163*
	1INS R	Insert	ATGCTGCATAGACCCAAAGG		264**
	1 HST R	Host right border	TAACCTTGTGCTTGCCATGA		
eTcBV1-II	2HST F	Host left border	CCACAAAAGATGAGAGAGAGACC	58	352*
	2INS R	Insert	GCCCAAATGGCATTACTAGCC		207**
	2HST R	Host right border	AGCAAAGTGCCCATTACCAC		
eTcBV1-III	3HST F	Host left border	TTCTGTTTGGCTGTACAGAGG	63	224*
	3INS F	Insert	ATCTCATGGGCCTTTCCTTT		299**
	3HST R	Host right border	GAAGTCAAAGGGGGAAAAGG		
TcAAR2	AAR2 F	Host right border	qRT-PCR	60	97
LOC18601928	AAR2 R	Host left border	ATGGGGCAATCACTTGAAGC		
			ACAGCTGACTCCTTGTACGG		
TcPAP26	PAP26 F	Host right border	AACACCTCCAGAGATTGGTCC	60	125
LOC18588841	PAP26 R	Host left border	ACAAGACAGTCTGTGCTCCAC		
TcACP1	ACP1 F	Left primer	CAGCGAGAAAAGTGCCTAGA	60	128
LOC18599903	ACP1 R	Right primer	AAATAAATAGACTTGAGTTCACAACAA		

*expected size without viral insertion.
**expected size with viral insertion.

found in the 3839 bp fragment with a 3680 bp overlap, and the right primer designed from the insert region found in the 4138 bp fragment with a 614 bp overlap. The complete assembled contig of the three clones comprised 10170 bp. A BLASTN search revealed 99% identity with B97 Chromosome IX of the contig sequence at the left (contig:1–224 nt, Chr IX:7512991–7513214 nt) and right (contig:9727–10170 nt, 7513215–7513658 nt) termini. Similarly, a BLASTN search for the 9502 bp inserted sequence found 87–100% identity for contig region 6059–6540 nt with multiple isolates of Cacao swollen shoot Ghana S virus. Pairwise distance analysis for the 9502 bp inserted sequence and 10 full-length cacao-badnaviral reference genomes in NCBI exhibited a 59.1 to 76.4 % nt identity among the 11 sequences. Notably, the eTcBV1-II sequence was found to be distinct, with maximum identity of 61.7% with CBSLV (Table 2).

In order to verify this assembly of the 10,170 nt contig formed from Sanger sequencing data of the amplicons from PA107 (eTcBV1-II), we generated an assembly from third party Illumina data (SRR9938225) generated from the same clone as a reference assembly. The 10,170 nt contig formed from Sanger sequencing data was used as the reference assembly. A scaffold of 10,204 nt was constructed (Extension of 13 and 20 nt on 5′ and 3′ ends into cacao genome) with 54.8X coverage depth (Fig. 2**d**). Both contigs shared 99% identity with no gap or indel in the common region.

Further analysis of the inserted sequence found in PA107 predicted six coding regions (≥100 codons) at nt 568–1173 (606 bp) for ORF1, 1173–1646 (474 bp) for ORF2, 1624–7311 (5686 bp) for ORF3, 3968–4348 (381 bp) for ORF4, 6371–6691 (321 bp) for ORF5 and 6939–7421 (483 bp) for ORFY, which encode proteins of 201, 157, 1895, 126, 106 and 160 aa, respectively. A BLASTP search revealed that these encoded proteins shared 23–36% aa sequence identity for ORF1, 25–37% for ORF2, and 36–58% for ORF3 with badnavirus proteins. No homology was found for the proteins encoded by ORFs 4, 5 and Y. A conserved domain search predicted a Domain of Unknown Function (DUF1319) in the protein encoded by ORF1 (85–182 aa). The protein encoded by ORF3, which showed 58% identity with the 1958 aa polyprotein from cacao yellow vein banding virus (GenBank accession YP_009345075), exhibited pepsin-like aspartate protease (Pep) reverse transcriptase (RT) and ribonuclease H (RNase H) domains at aa 1120–1210, 1356–1541 and 1637–1765, respectively. No conserved domains were identified in ORFs 2, 5, X and Y.

In order to compare the features predicted in the eTcBV1-II, we analysed genome sequences of 87 cacao-infecting badnavirus isolates/species in NCBI. The total length of these genome sequences ranged from 6820 bp for cacao red vein virus (CRVV) isolate NIG18 (GenBank accession MH029282) to 7533 bp for cacao yellow vein banding virus (CYVBV) isolate SCA6 (KX276640) (**Supplementary Table S4**). Unfortunately, the criteria used to predict ORFs and conserved domains (CD) are not consistent between the various previous studies (Chingandu et al., 2017; Muller et al., 2018). For uniformity, we used a minimum ORF length ≥ 100 codons in prediction of ORFs and did not consider incomplete CD found in predicted ORFs. Using this criterion, the search predicted four ORFs (ORFs 1–3 and Y) in all genomes. Eleven genomes contained an additional ORF (ORFX:10, ORF4:1). Six ORFs (ORFs 1–4, X and Y) were predicted in cacao swollen shoot CD virus (CSSCDV) isolate Buyo2 (GenBank accession MN433935), whereas cacao swollen shot Ghana R virus (CSSGRV) isolate Gha39-15 (GenBank accession MF642733) contained ORFs 1, 3 and Y only (**Supplementary Table S4**). The ORF1 of the 87 badnaviruses varied from 417 to 522 bp, whereas the ORF2 ranged from 297 to 450 bp. The ORF3, which is the longest ORF, ranged from 4779 to 5877 bp. The ORFY contained 348 to 441 bp, whereas the ORFX, found in 11 genomes, ranged from 309 to 441 bp. CSSV isolate CI569-10 (GenBank accession MF642717) and CSSCDV isolate Buyo2 contained ORF4 of 309

Table 2
Similarity matrix based on sequence of reference genomes of 10 cacao-infecting species in the genus Badnavirus, together with the 9502 bp inserted sequence of endogenous *T. cacao* bacilliform virus1-II (eTcBV1-II) identified in this study. The 10 species comprise:- Cacao swollen shoot Togo A virus, AJ781003.1 (CSSTAV); Cacao swollen shoot Togo B virus, MN179743 (CSSTBV); Cacao swollen shoot CD virus, NC_038378 (CSSVCD); Cacao swollen shoot Ghana M virus NC_043534 (CSSGMV); Cacao swollen shoot Ghana N virus, NC_040622 (CSSGNV); Cacao swollen shoot CE virus, NC_040692 (CSSVCE); Cacao swollen shoot Ghana Q virus, NC_043535 (CSSGQV); Cacao mild mosaic virus, NC_033738 (CaMMV); Cacao bacilliform Sri Lanka virus, NC_040809 (CBSLV); Cacao yellow vein banding virus, NC_033739 (CYVBV).

Accession	1	2	3	4	5	6	7	8	9	10	11
1-CSSTAV	100.0										
2-CSSTBV	76.4	100.0									
3-CSSCDV	75.8	75.0	100.0								
4-CSSGMV	72.2	73.1	71.3	100.0							
5-CSSGNV	70.2	70.8	70.4	75.4	100.0						
6-CSSCEV	70.9	71.8	70.3	70.1	69.2	100.0					
7-CSSGQV	62.4	62.0	62.3	63.9	62.9	62.8	100.0				
8-CaMMV	61.6	60.8	60.7	61.7	61.3	61.3	60.9	100.0			
9-CBSLV	60.6	60.9	60.6	60.1	60.4	61.5	61.1	60.4	100.0		
10-eTcBV1-II	60.5	59.4	59.9	59.7	61.1	61.0	60.0	60.5	61.7	100.0	
11-CYVBV	60.5	60.5	60.8	60.3	60.7	59.1	60.1	60.0	60.5	60.3	100.0

and 324 bp, respectively. The inserted eTcBV1-II, which we identified in PA107, agrees with these findings in terms of number, size, and arrangement of the ORFs, though the ORFs are slightly longer (**Supplementary Table S4, S5**). The additional ORF, designated as ORF5, found in eTcBV1-II is unique and did not exist in the cacao badnavirus species analysed (**Supplementary** Fig. 2).

A conserved domain search identified the DUF1319 domain in the protein encoded by ORF1 in all 87 genomes. A trimeric autotransporter adhesin (TAA), the Trp ring domain was found in the ORF2-encoded protein in CSSV isolates GH64, Gha53-15 and Gha57-15 (GenBank accessions KX592572, NC_040693, NC_04353). Three to five conserved domains were identified in the polyprotein encoded by ORF3. Pep, RT and RNase H domains were predicted in all genomes, except CSSV isolate CIS3 (GenBank accession KX592576, which lacked the RNase H domain. The Zinc binding motif, Zinc knuckle, was also identified in all genomes, except CSSV isolate Buyo15, CSSCDV isolate CI152-09, CSSCEV isolate CI632-10, CSSGQV isolate Gha2-15 and new world CaMMV isolate PR3 (MN433938, NC_038378, MF642719, MF642726, MW052520). The universal minicircle sequence binding protein (UMSBP) domain was found in 11 genomes (**Supplementary Table S5**). No conserved domains were identified in ORFs 4, X and Y (**Supplementary Table S6**). The number and types of conserved domains identified in eTcBV1-II correspond to the domains predicted in CSSV and cacao infecting viruses discovered in the Americas (**Supplementary Tables S4, S5**).

The phylogenetic tree based on complete sequences of 41 selected genomes representing 10 cacao badnavirus formed four major clusters. West African cacao badnaviruses, grouped in a distinct cluster A, except CSSGQV which formed cluster B. The CYVBV and CBSLV, along with eTcBV1-II, appear to be basal to the clusters formed by West African cacao badnaviruses and were placed in cluster C. The cluster D contained four isolates of CaMMV (Fig. 3).

A tree was also constructed using the amino acid sequence of ORF3 from 55 out of 59 species classified in genus Badnavirus in ICTV, together with eTcBV1-II, to ascertain phylogenetic relationships (**Supplementary Fig. S3**). West African cacao badnaviruses grouped together in a major subcluster of the main cluster 1 with the exception of CSSGQV, which was found to be more similar to Mulberry badnavirus 1. The second subcluster is mainly formed by bacilliform virus (BV) species discovered in Dioscorea (yam) and Taro. The second main cluster is formed mostly by Banana streak virus species and badnavirus species identified in sugarcane, along with the species found in some ornamental plants. The CYVBV, CBSLV and eTcBV1-II sequences did not show close similarity to the badnavirus species that cause major yield losses.

3.2.2. eTcBV1-I

A 7052 bp contig from GU123/V assembly aligned from position 2692–4820 nt with eTcBV1-I. A region of the contig from position 1–2274 nt showed 98% identity with Chromosome VII region 2625436–2627707 nt (GenBank accession NC_030856). The insertion site of the putative viral sequence is located in exon 3 of the *TcRGA3* gene, which encodes the disease resistance protein RGA3 (GenBank accession LOC18593610) (Fig. 1**a**).

The primers designed from cacao genomic regions bordering the insert successfully amplified the expected size band from GU123/V (Reading accession RUQ 1068), which was cloned and sequenced. The sequencing data confirmed the insertion site identified in the *de novo* assembly. The complete assembled contig of the clone comprised 5699 bp, with the sequences being 99% identical with B97 Chromosome VII at the left (2627655–2627707 nt) and right (2628354–2627708 nt) termini, respectively, and harbouring an insertion of 4999 bp (Fig. 2**a**).

This insert contained coding regions at 4137–4706 nt for ORF1, 3664–4137 for ORF2, 54–3686 for ORF3 and 959–1339 for ORF4, which encode proteins of 189, 157, 1211 and 126 aa, respectively (minus strand). Notably, the contig lacks ORF5 and ORF Y. The arrangement of the ORFs revealed an inverted integration of eTcBV1-I in the host genome. The locations of the predicted ORFs and conserved domains are similar to those found in eTcBV1-II. The ORF1 and ORF3 are shorter in length compared to eTcBV1-II, with the differences probably due to deletions that occurred during the process of integration into the cacao genome. A conserved domains search predicted the DUF1319 domain in ORF1. The ORF3 contained Pep domain only as the part of the polyprotein towards the C terminus that contained RT and RNase H region is truncated (Fig. 2**a**).

Alignment of the sequences of proteins predicted in ORFs 1–4 of eTcBV1-1 revealed 87, 92, 94 and 85% identity with the proteins predicted in eTcBV1-II (**Supplementary Fig. S4, S5**).

3.2.3. eTcBV1-III

Reference eTcBV1-III sequence was mapped to a 3138 bp contig constructed from NA33 data with contig position from 975 to 2122 nt (plus/minus). The contig region from position 1–609 nt was 99% identical with the Chromosome IV region 20394613–20395221 nt (GenBank accession NC_030853). A putative insertion site of eTcBV1-III was found in intron 8 of the *TcAAR2* gene, which encodes a protein AAR2 homolog (GenBank accession LOC18601928) (Fig. 1**c**).

The primers designed from cacao genomic regions bordering this insert did not amplify any band from NA33 (Reading accession

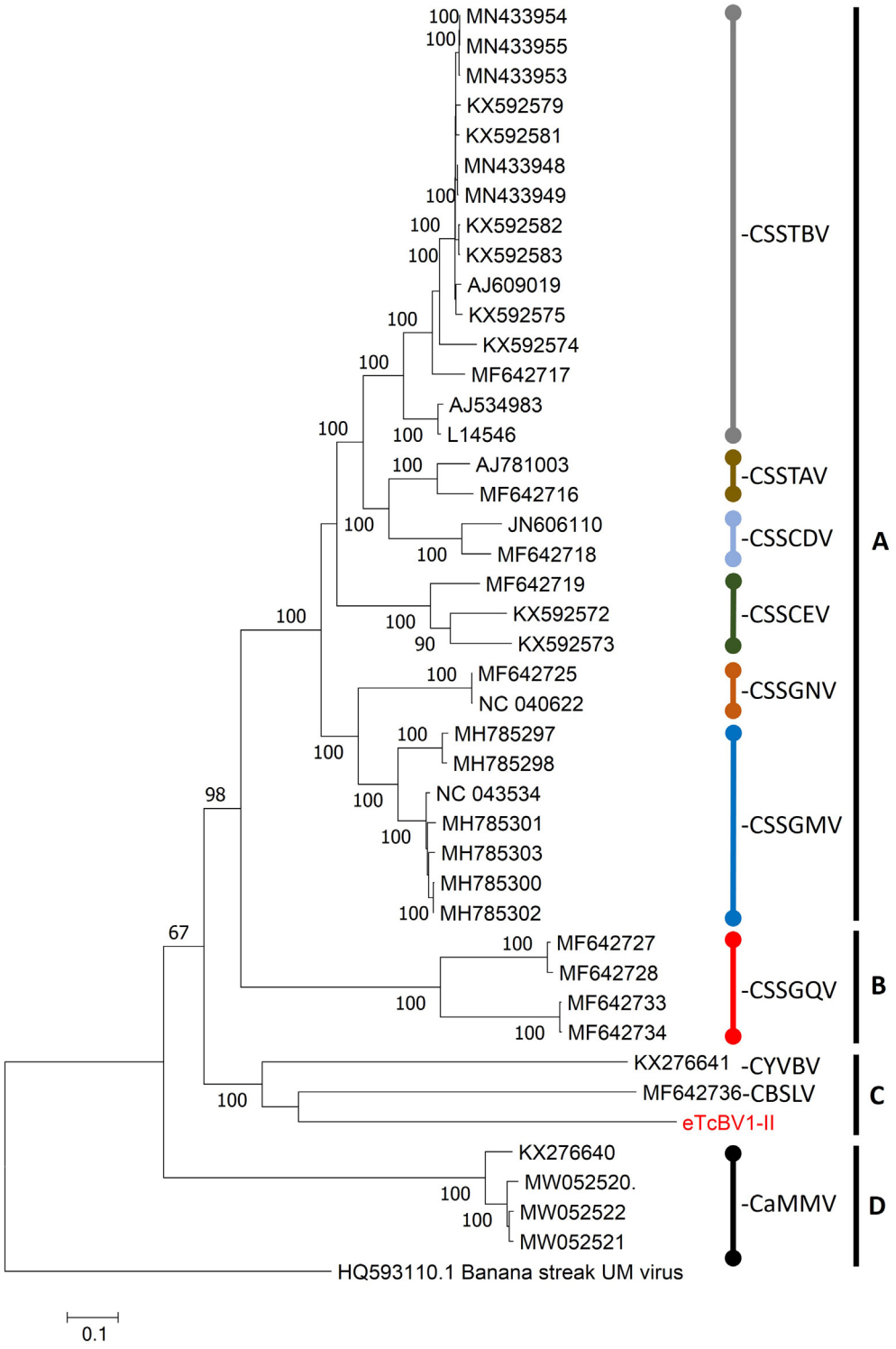

Fig. 3. Maximum Likelihood tree, generated in MEGA11, based on complete nucleotide sequence of 41 selected genomes representing 10 cacao-infecting species in genus Badnavirus and the 9502 bp inserted sequence of endogenous *T. cacao* bacilliform virus1-II (eTcBV1-II) identified in this study. The bootstrap support values are shown above the main tree nodes. The ten badnavirus species comprise Cacao mild mosaic virus (CaMMV); Cacao bacilliform Sri Lanka virus (CBSLV); Cacao yellow vein banding virus (CYVBV); Cacao swollen shoot Ghana Q virus (CSSGQV); Cacao swollen shoot Ghana M virus (CSSGMV); Cacao swollen shoot Ghana N virus (CSSGNV); Cacao swollen shoot CE virus (CSSCEV); Cacao swollen shoot CD virus (CSSCDV); Cacao swollen shoot Togo A virus (CSSTAV) and Cacao swollen shoot Togo B virus (CSSTBV).

RUQ 1577). Therefore, two primer sets were designed from left and right border host sequence and insert with considerable overlap.

The PCR with these primers amplified two bands of around 4.5 and 1.0 kb. Assembly of the sequencing data obtained from both bands surprisingly formed two independent contigs. The alignment of both contigs revealed a direct repeat of 282 nt located in the insert region used to design the primers, a finding that explains amplification of two partial sequences instead of the targeted full insert. The repeats are located at 4155–4436 in contig1 and 1–282 nt in contig2 (**Supplementary Fig. S6**). BLASTN search of the

large contig of 4651 bp found 99% identity with B97 Chromosome IV (20395075–20395221 nt) at the left terminus. The right terminus of the second 1098 bp contig showed 99% identity with B97 Chromosome (20395241–20395317 nt). These results confirm the insertion site found in the *de novo* assembly; however, there was a deletion of 19 nt of the host genome chromosome IV from 20395222 to 40 nt (Fig. 2**c**).

Analysis of both contigs revealed an inverted integration of eTcBV1-III in the host genome. The insert sequence found in contig1 contained coding regions at 848–4558 nt for ORF3, 3802–4182 for ORF4, 1468–1788 for ORF5, and 738–1220 for ORFY, which encode proteins of 1236, 126, 106 and 160 aa, respectively (minus strand). A partial ORF2 that encodes a 101 aa protein was predicted in contig 2 at 232 to 537 nt. Similar to the eTcBV1-II, the arrangement of the ORFs revealed an inverted integration of eTcBV1-III in the host genome. The locations of the predicted ORFs and conserved domains are similar to those found in eTcBV1-II; however, it lacks ORF1. Additionally, the ORF2 and ORF3 are shorter in length compared to those in eTcBV1-II. Also, the polyprotein encoded by the ORF3 is truncated, as around 600 aa are deleted from N terminus; however, the three conserved domains comprising Pep RT and RNase H region are intact (Fig. 2**c**).

Proteins predicted in ORFs 2–5 and Y in eTcBV1II-1 shared 84, 98, 95, 93 and 98% identity with the respective proteins predicted in eTcBV1-II (**Supplementary Fig. S4, S5**).

3.3. Screening of germplasm

A multiplex assay was developed to screen the 342 germplasm accessions held at ICQC-R for identification of type of eTcBV1s, and zygosity status of the identified eTcBV1 loci. The details on design of the assay are provided in Fig. 4a and Table 1. This screening identified eTcBV1s I-III in 103 accessions (**Supplementary Table S7**). The eTcBV1-I locus was found in 30 accessions, all in a hemizygous status. The eTcBV1-II locus was identified in homozygous and hemizygous form in 20 and 26 accessions, respectively. The eTcBV1-III locus and eTcBV1-III were discovered in homozygous and hemizygous form in 6 and 39 accessions, respectively. Among the 103 accessions, 16, 27 and 41 accessions independently contained the eTcBV1s I, II and III, respectively. The combination of the eTcBV1s I and II was found in 15 accessions, whereas four accessions contained combination of the eTcBV1s II and III (Fig. 4**b**).

3.4. Inheritance study

Selfed progenies of CRU12, which contained both eTcBV1-I and II, and NA702, which contained eTcBV1-III, were screened with the multiplex PCR assay to assess the inheritance pattern of the insert. The status of the eTcBV1s loci were hemizygous in both accessions. The progenies of both accessions segregated in the classic Mendelian ratio of 1:2:1 (Fig. 4**c**), a result that confirms the endogenized nature of the sequences.

3.5. Effect of eTcBV1 on expression of hosting genes

The eTcBV1s identified in this study are located in the coding region of cacao genome. Therefore, we analysed expression of cacao *TcRGA3, TcPAP26* and *TcAAR2* in subsets of cacao germplasm. We selected eight cacao accessions found free of the eTcBV1s reported in this study, and four accessions that each contain eTcBV1-I, II or III. Third party RNA-seq datasets from BioProject PRJNA558793 (Hämälä et al., 2021) were used for transcriptome analysis. We did not find transcripts for *TcRGA3* gene in any accession with or without eTcBV1s. The *TcAAR2* gene expressed 1.1 to 2.4 times higher compared to *TcPAP26* in seven out of eight accessions found free of eTcBV1s (Fig. 5**a**). A similar trend was found in the accessions that contain eTcBV1-II in *TcPAP26* (Fig. 5**c**). Similarly, the eTcBV1-III insertion in the *TcAAR2* gene has significantly reduced expression of the gene, which was 1.3 to 5.4 times lower compared to *TcPAP26*, in four accessions that contain this insertion (Fig. 5**b**). The effect of insertion on expression of *TcAAR2* was more profound in NA916 and NA710 and NA807 (Fig. 5**b**).

For precise quantitation, we conducted qRT-PCR to determine change in expression level of both genes in leaves of six selected cacao accessions, three of which have eTcBV1-III insertion in the *TcAAR2* gene whereas the other three have an eTcBV1-II insertion in the *TcPAP26* gene. The expression of *TcAAR2* was significantly lower in NA916, NA710, NA232 (Reading accessions RUQ 1334, 1588, 1504, respectively) that contain eTcBV1-III insertion compared to the eTcBV1-III insertion free PA299, PA107 and PA70 (Reading accessions RUQ 1621, 1728 and 33, respectively). Similar results were found for the impact of eTcBV1-II in *TcPAP26* on expression pattern. Relative quantitation revealed significant reduction in expression of *TcPAP26* in PA70, PA107 and PA299 that contain an eTcBV1-II insertion compared to insertion free accessions (Fig. 5**d**).

3.6. Phylogeny cacao accession

Data on status of cacao accessions held at ICQC-R for four types of eTcBV1s as obtained from this study and our previous work, and availability of single nucleotide polymorphism (SNP) marker fingerprints provided an opportunity to generate a phylogenetic tree and map the types of viral insertions. Clustering of the cacao accessions in genetic groups, based on SNP markers, was found to be correlated with the type of insert. For example, a single eTcBV1-I insertion was found in the Guiana accessions, which formed a distinct cluster. Another cluster, comprising accessions from the Marañon genetic group, had an eTcBV1-II insertion, either alone or in combination with an eTcBV1-I insertion. One distinct cluster, subdivided into two sub-clusters comprised the Trinitario and admixed accessions that have an eTcBV1-VI insertion. All reference accessions in the Nanay cluster contained an eTcBV1-III insertion (Cornejo et al., 2018). A distinct group formed by PNG accessions contain eTcBV1-III alone or in combination with an eTcBV1-VI insertion. These PNG accessions are cloned hybrids which were developed in Papua New Guinea through direct selection from F_1 progenies of Trinitario × Amazon crosses and are considered to have resistance against *Phytophthora* pod rot (https://www.cocoaboard.org.pg/wp-content/uploads/2019/09/PNG_ExtensionManual_final_draft25Aug17.pdf). Though exact information on the name of parental clones is not available, it can be speculated that the hybrids may have inherited the insertions from Trinitario and Amazonian (possibly Nanay) parents, that have eTcBV1-III and VI insertions, respectively. The GEBP accessions are also hybrids developed under the Genetic Enhancement for Black Pod programme. These accessions did not form a specific cluster, a result that could be expected considering the diverse nature of their parents (**Supplementary Table S7**). The data on insertion type present in the parent is lacking, except for GEBP 565 (PA124 × IMC103) GEBP 571 (PA124 × IMC103). Both accessions have eTcBV1-III which is also present in parental line PA124 (Fig. 6).

4. Discussion

We have previously demonstrated the presence of a diverse array of endogenous badnaviruses in cacao genome and proposed to name these integrations *eTcBV1* and *eTcBV2* for endogenous *T. cacao* bacilliform virus 1 (species S) and 2 (Species S prime), respectively. We also reported comprehensive analysis of one

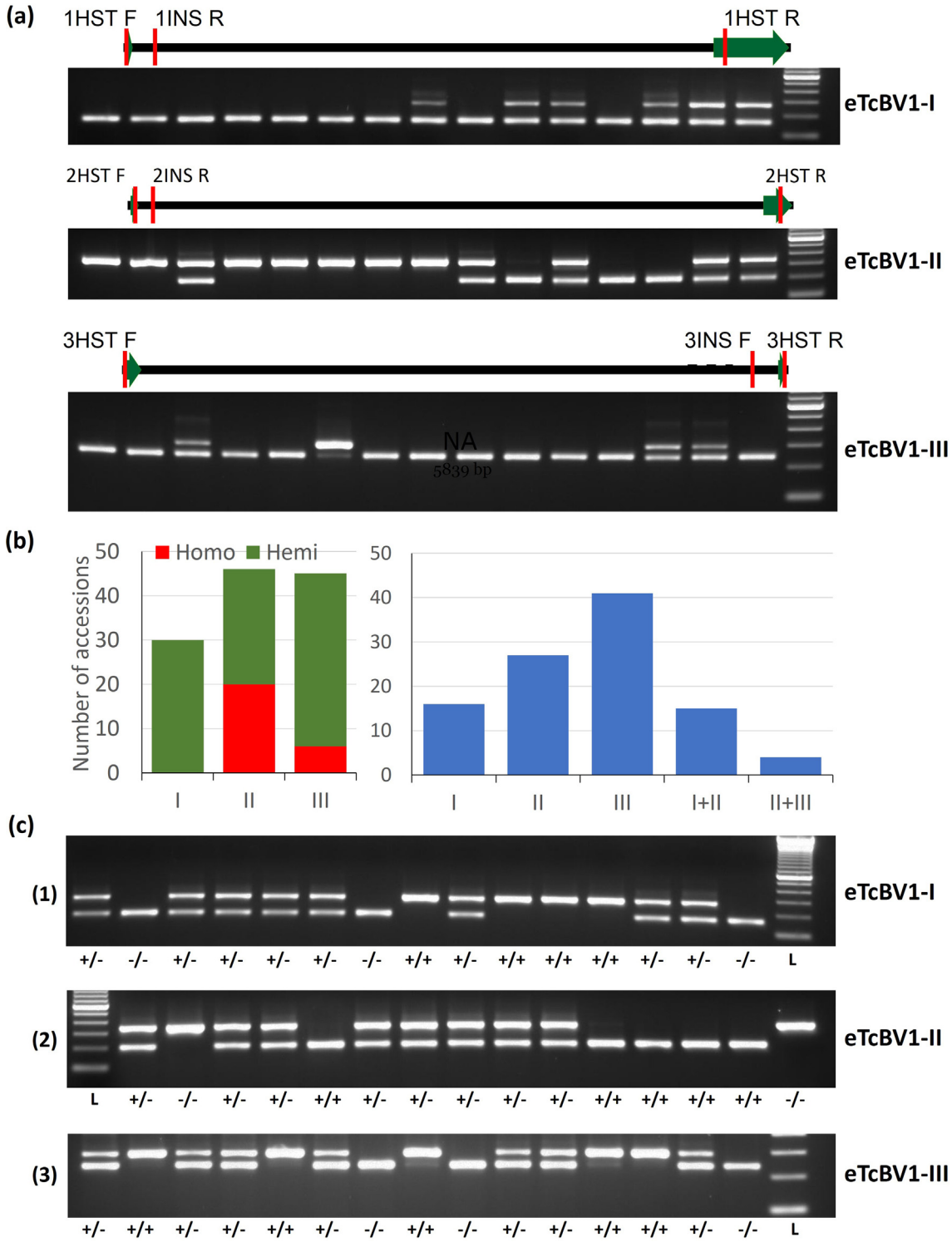

Fig. 4. Prevalence and inheritance of endogenous *T. cacao* bacilliform virus1 (eTcBV1) in cacao. (a) Green arrows on both sides represent genomic region of cacao (B97 *T. cacao* reference genome) bordering an eTcBV1s. Black horizontal lines represent inserted eTcBV1 region. Vertical red lines denote location of primers. Genotyping of a selected subset of the germplasm is shown in gel images. (b) Green and red bars represent number of accessions with a specific eTcBV1 in homozygous and heterozygous status, respectively. Blue bars represent number of accessions that contain single or multiple eTcBV1. (C) Genotyping of selfed progeny of ARF12 (eTcBV1s I and II) and NA702 (eTcBV1-III). The signs "−/−", "+/+" and "+/−" under each lane represents an allele lacking viral insertion, or homozygous and hemizygous status of the eTcBV1 type, respectively.

specific type eTcBV1-VI insert in the PA 279 cacao clone. Here we report analysis of three additional eTcBV1 integrants in the genome of different cacao clones.

Most of the studies in this area either use bioinformatic tools or PCR approach to identify EVEs. However, the sequences identified by such approaches may have been derived from episomal or integrated viral DNA, particularly as many databases are incorrectly annotated. We have developed a more practical and precise approach. Initially, we screened the cacao genomic sequence data to detect the putative EVE sequences, then used *de novo* assemblies

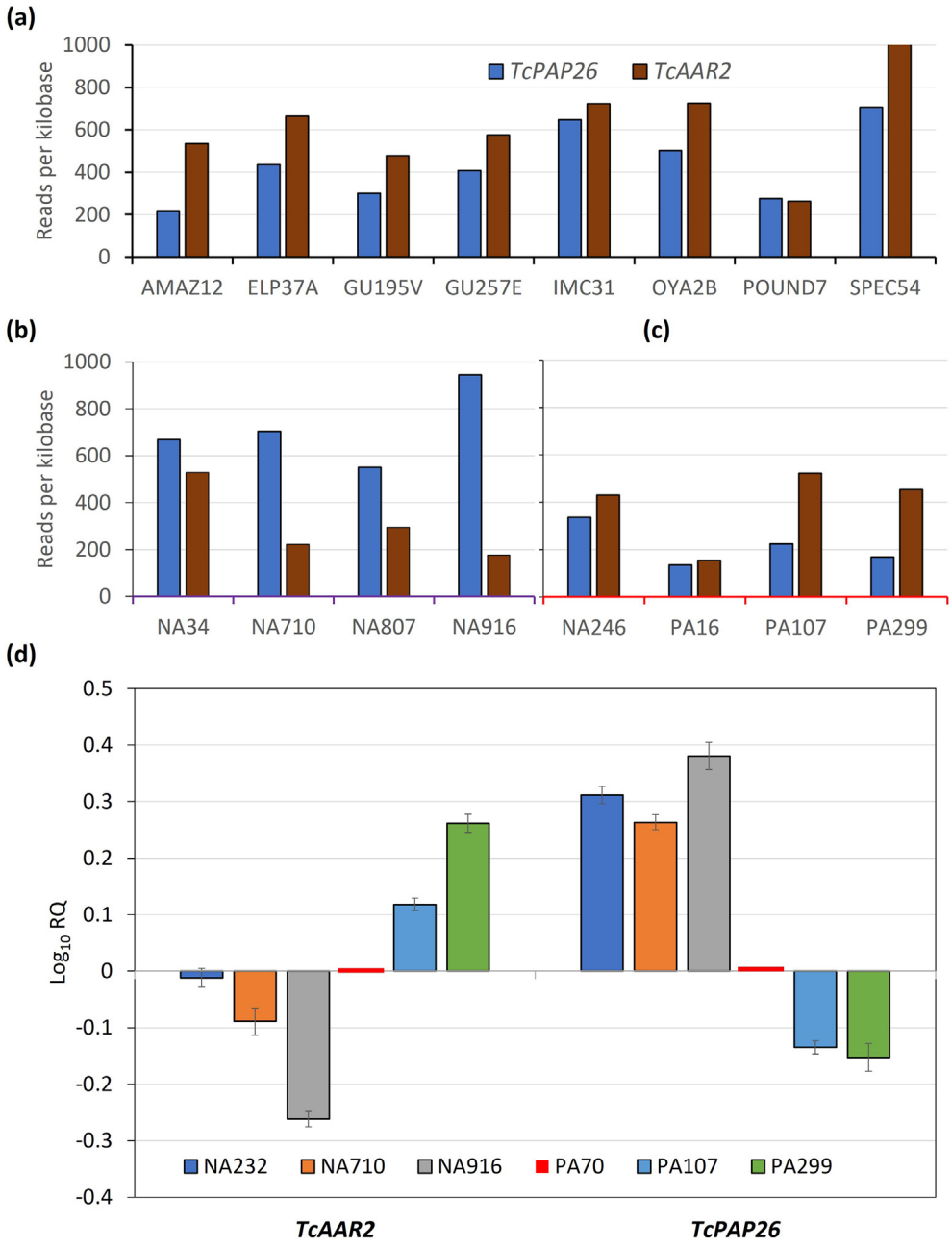

Fig. 5. Effect of insertions on gene expression. (a) Co-expression of the cacao genes that encode bifunctional purple acid phosphatase 26 (TcPAP26; LOC18588841) and AAR2 homolog proteins (TcAAR2; LOC18601928) in subsets of cacao accessions found free of the endogenous T. cacao bacilliform virus1 (eTcBV1) described in this study. (b) Accessions that contain eTcBV1-III in TcAAR2 gene. (c Accessions that contain eTcBV1-II in TcPAP26 gene. (d) Effect of insertion on expression of TcAAR2 and TcPAP26 genes, as measured by qRT-PCR, in six selected cacao accessions. TcACP1 gene (LOC18599903) was used as a reference gene to normalize expression level. Third party RNA-seq datasets from BioProject PRJNA558793 (Hämälä et al., 2021) were used for this co-expression analysis.

to find the putative insertion site in the cacao genome, amplified endogenized sequences along with the bordering cacao sequence with the primers designed from the cacao genome, and sequenced them for identification of the sequence of EVEs and precise location of the insert in the host genome. Importantly, we also conducted an inheritance study to prove that the insertions are inherited as cacao loci. This comprehensive approach eliminates the possibility of detection of episomal DNA as EVEs.

As regards the overall background of these inserts in relation to the genetic diversification of cacao, the previous study (Muller et al., 2021) provided initial evidence based on the presence of inserts described. This analysis has been significantly extended in the present study, first by providing a more detailed phylogeny (Fig. 6) and also by examining the effect of the insert on the transcription of the host gene. This latter feature has never previously been studied in any crop, although it is a fundamental aspect of the discussion around whether and how the insert is maintained within the genome. If there was no selective advantage, then it would be expected that the insert would gradually be lost by a process of mutation leading to inactivation. The fact that a range of inserts have been maintained, without the introduction of stop codons, during the diversification of cacao, and that they positively or negatively affect transcription of the host gene, suggests instead that they play a functional role in the growth and development of

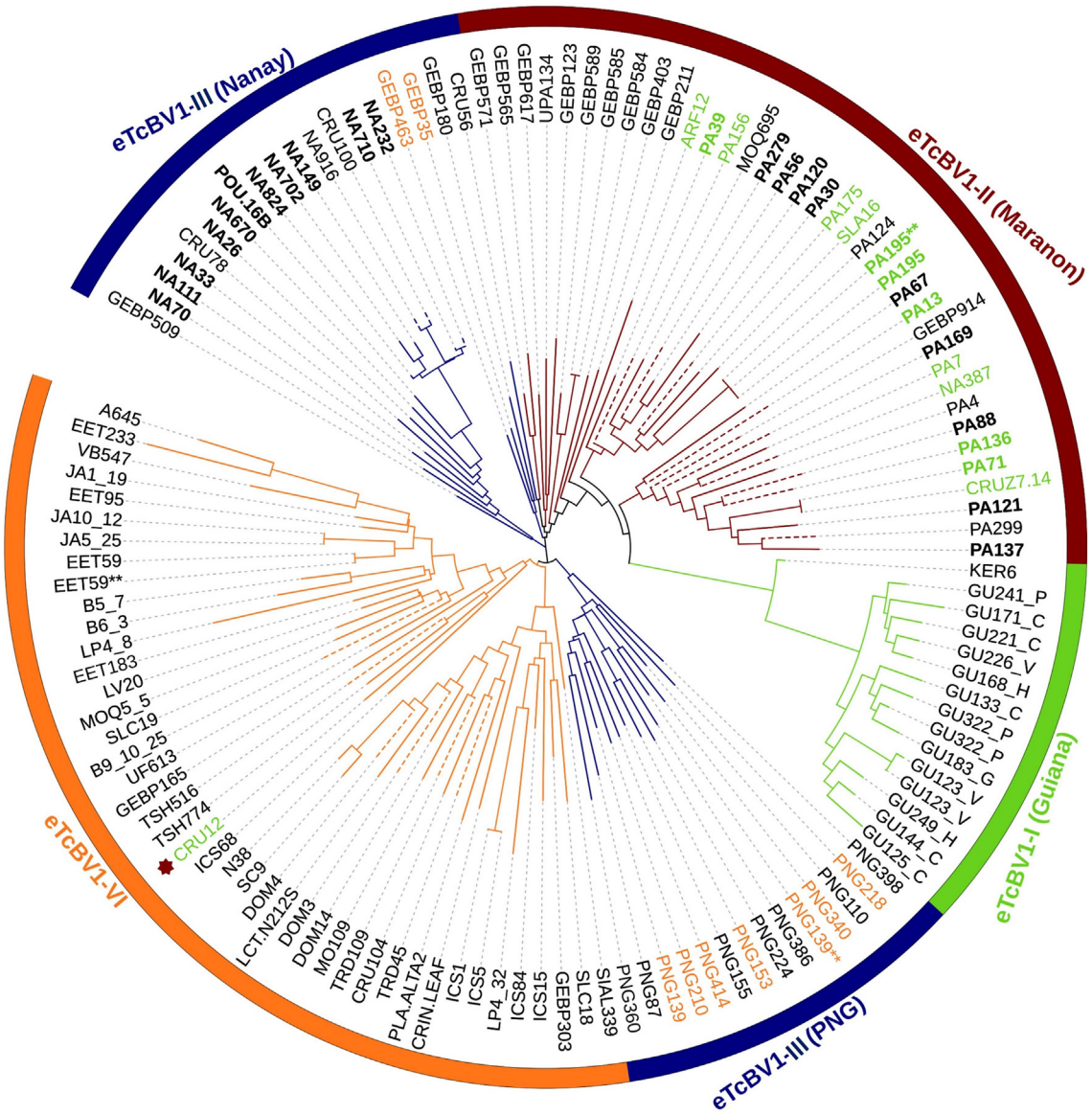

Fig. 6. A phylogenetic tree of cacao accessions. Maximum Likelihood tree based on distance matrix generated in MEGA11 using 151 SNP markers for 128 cacao accessions, from ICQC-R, which are found positive for endogenous *T. cacao* bacilliform virus1 (eTcBV1). Node and strip colour denote a specific type of eTcBV1 identified:- Green:eTcBV1-I; Maroon:eTcBV1-II; Navy blue:eTcBV1-V; Brown:eTcBV1-VI. Label font colours other than black represent accessions with multiple types of eTcBV1. Bold font of a label denotes reference accessions representing a particular genetic group. The single accession that contains three types of inserts is marked with a star.

the crop. Analysis of possible non-coding RNAs produced by the inserts, and a possible protection against further viral infection, is another obvious area for research. In this latter context, another notable finding from the present study is that the codon usage of the inserts remains similar to that of the invasive virus and has not been homogenized to become more similar to that of the host genome and therefore more plant-like (**Supplementary Table S8, 9**).

The findings reported here also have other specific and more general implications in the context of genome manipulation and evolution. In the former area, the fact that these inserts have stably survived over a long period during cacao domestication and genetic diversification suggests that the insert sites may represent possible "safe landing sites" for transgene insertion (Dong and Ronald 2021). At a more general level, the continuing increase in the number of viral inserts identified in cacao and other crops adds to the understanding of the dynamic and plastic nature of the genome. Such findings, and the possibility of interaction between integrated and episomal forms, should also be considered in the context of extrachromosomal circular DNA (eccDNA), a type of DNA that exists in many eukaryotes including plants (Wang et al., 2021). These latter authors identified 743 eccDNAs in Arabidopsis, with a frequency that was tissue-specific. Another relevant recent finding is the identification of such eccDNAs in Palmer amaranth and their amplification during the rapid evolution of glyphosate resistant biotype (Spier Camposano et al., 2022). A further level of complexity is suggested by the finding that this latter form of eccDNAs can be tethered to the end of chromosomes (Koo et al., 2021). To date, there is no information about the presence of eccDNAs in cacao.

In summary, the present study has significantly extended our knowledge of endogenous badnaviral sequences in cacao and suggested several areas for further investigation.

5. Conclusion

Overall, this study represents a significant extension to the understanding of badnavirus sequences integrated into the genome of cacao. From a germplasm management perspective, it reinforces the need to extend the search for additional integrated sequences as a means to discriminate between episomal and integrated sequences. Also, in terms of the retention of such integrated sequences during evolution and diversification of cacao germplasm, it suggests a possible functional role, namely in affecting the transcription of the host gene.

Funding

This research was funded by the Cocoa Research Association Ltd (CRA), Cocoa Research UK Ltd, and the United States Department of Agriculture (USDA).

Data availability statement

Sequencing data generated in this study were submitted to NCBI GenBank and have accession numbers OP351634-OP351637 and Third Party Annotation (TPA) database accession number BK059625.

CRediT authorship contribution statement

Ihsan Ullah: Conceptualization, Methodology, Resources, Software, Writing – original draft, Writing – review & editing, Validation, Formal analysis, Investigation. **Jim M. Dunwell:** Conceptualization, Resources, Validation, Writing – review & editing, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.sjbs.2023.103648.

References

Bhat, A.I., Mohandas, A., Sreenayana, B., Archana, T.S., Jasna, K., 2022. Piper DNA virus 1 and 2 are endogenous pararetroviruses integrated into chromosomes of black pepper (*Piper nigrum* L). VirusDisease 33 (1), 114–118. https://doi.org/10.1007/s13337-021-00752-w.

Blum, M., Chang, H.Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., Raj, S., Richardson, L., Salazar, G.A., Williams, L., Bork, P., Bridge, A., Gough, J., Haft, D.H., Letunic, I., Marchler-Bauer, A., Mi, H., Natale, D.A., Necci, M., Orengo, C.A., Pandurangan, A.P., Rivoire, C., Sigrist, C.J.A., Sillitoe, I., Thanki, N., Thomas, P.D., Tosatto, S.C.E., Wu, C.H., Bateman, A., Finn, R. D., 2021. The InterPro protein families and domains database: 20 years on. Nucleic Acids Res. 49 (D1), D344–D354. https://doi.org/10.1093/nar/gkaa977.

Boutanaev, A.M., Nemchinov, L.G., 2021. Genome-wide identification of endogenous viral sequences in alfalfa (*Medicago sativa* L.). Virol. J. 18 (1), 185. https://doi.org/10.1186/s12985-021-01650-9.

Brown, J.R., 2003. Ancient horizontal gene transfer. Nat. Rev. Genet. 4 (2), 121–132. https://doi.org/10.1038/nrg1000.

Chingandu, N., Zia-Ur-Rehman, M., Sreenivasan, T.N., Surujdeo-Maharaj, S., Umaharan, P., Gutierrez, O.A., Brown, J.K., 2017. Molecular characterization of previously elusive badnaviruses associated with symptomatic cacao in the new world. Arch. Virol. 162 (5), 1363–1371. https://doi.org/10.1007/s00705-017-3235-2.

Cornejo, O.E., Yee, M.-C., Dominguez, V., Andrews, M., Sockell, A., Strandberg, E., Livingstone, D., Stack, C., Romero, A., Umaharan, P., Royaert, S., Tawari, N.R., Ng, P., Gutierrez, O., Phillips, W., Mockaitis, K., Bustamante, C.D., Motamayor, J.C., 2018. population genomic analyses of the chocolate tree, *Theobroma cacao* L., provide insights into its domestication process. Commun. Biol. 1, 167. https://doi.org/10.1038/s42003-018-0168-6.

Dong, O.X., Ronald, P.C., 2021. Targeted DNA insertion in plants. Proc. Natl. Acad. Sci. U. S. A. 118 (22), e2004834117. https://doi.org/10.1073/pnas.2004834117.

Gayral, P., Noa-Carrazana, J.-C., Lescot, M., Lheureux, F., Lockhart, B.E.L., Matsumoto, T., Piffanelli, P., Iskra-Caruana, M.L., 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.

Gregor, W., Mette, M.F., Staginnus, C., Matzke, M.A., Matzke, A.J.M., 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.

Hämälä, T., Wafula, E.K., Guiltinan, M.J., Ralph, P.E., dePamphilis, C.W., Tiffin, P., 2021. Genomic structural variants constrain and facilitate adaptation in natural populations of *Theobroma cacao*, the chocolate tree. Proc. Natl. Acad. Sci. 118 (35), e2102914118. https://doi.org/10.1073/pnas.2102914118.

Johnson, W.E., 2019. Origins and evolutionary consequences of ancient endogenous retroviruses. Nat. Rev. Microbiol. 17 (6), 355–370. https://doi.org/10.1038/s41579-019-0189-2.

Kandito, A., Hartono, S., Trisyono, Y.A., Somowiyarjo, S., 2022. First report of cacao mild mosaic virus associated with cacao mosaic disease in Indonesia. New Dis. Reports 45 (2), e12071.

Katzourakis, A., Gifford, R.J., 2010. Endogenous viral elements in animal genomes. PLOS Genet. 6 (11), 1–14. https://doi.org/10.1371/journal.pgen.1001191.

Koo, D.H., Sathishraj, R., Friebe, B., Gill, B.S., 2021. Deciphering the mechanism of glyphosate resistance in *Amaranthus palmeri* by cytogenomics. Cytogenet. Genome Res. 161 (12), 578–584. https://doi.org/10.1159/000521409.

Langmead, B., Salzberg, S., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9 (4), 357–359. https://doi.org/10.1038/nmeth.1923.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25 (16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

Li, Y., Zhang, G., Cui, J., 2022. Origin and deep evolution of human endogenous retroviruses in pan-primates. Viruses 14 (7), 1370. https://doi.org/10.3390/v14071370.

Lockhart, B.E., Menke, J., Dahal, G., Olszewski, N.E., 2000. Characterization and genomic analysis of tobacco vein clearing virus, a plant pararetrovirus that is transmitted vertically and related to sequences integrated in the host genome. J. Gen. Virol. 81 (6), 1579–1585. https://doi.org/10.1099/0022-1317-81-6-1579.

Lopez-Gomollon, S., Müller, S.Y., Baulcombe, D.C., 2022. Interspecific hybridization in tomato influences endogenous viral sRNAs and alters gene expression. Genome Biol. 23 (1), 120. https://doi.org/10.1186/s13059-022-02685-z.

Motamayor, J.C., Lachenaud, P., da Silva, E., Mota, J.W., Loor, R., Kuhn, D.N., Brown, J. S., Schnell, R.J., 2008. Geographic and genetic population differentiation of the Amazonian chocolate tree (*Theobroma cacao* L). PLoS One 3 (10), e3311.

Muller, E., Ravel, S., Agret, C., Abrokwah, F., Dzahini-Obiatey, H., Galyuon, I., Kouakou, K., Jeyaseelan, E.C., Allainguillaume, J., Wetten, A., 2018. Next generation sequencing elucidates cacao badnavirus diversity and reveals the existence of more than ten viral species. Virus Res. 244, 235–251. https://doi.org/10.1016/j.virusres.2017.11.019.

Muller, E., Ullah, I., Dunwell, J.M., Daymond, A.J., Richardson, M., Allainguillaume, J., Wetten, A., 2021. Identification and distribution of novel badnaviral sequences integrated in the genome of cacao (*Theobroma Cacao*). Sci. Rep. 11 (1), 8270. https://doi.org/10.1038/s41598-021-87690-1.

Ndowora, T., Dahal, G., LaFleur, D., Harper, G., Hull, R., Olszewski, N.E., Lockhart, B., 1999. Evidence that badnavirus infection in *Musa* can originate from integrated pararetroviral sequences. Virology 255 (2), 214–220. https://doi.org/10.1006/viro.1998.9582.

Osorio-Guarín, J.A., Berdugo-Cely, J.A., Coronado-Silva, R.A., Baez, E., Jaimes, Y., Yockteng, R., 2020. Genome-wide association study reveals novel candidate genes associated with productivity and disease resistance to *Moniliophthora* spp. in cacao (*Theobroma cacao* L.). G3 (Bethesda). 10 (5), 1713–1725. https://doi.org/10.1534/g3.120.401153

Pfeiffer, P., Hohn, T., 1983. Involvement of reverse transcription in the replication of cauliflower mosaic virus: a detailed model and test of some aspects. Cell 33 (3), 781–789. https://doi.org/10.1016/0092-8674(83)90020-X.

Puig, A.S., 2021. Detection of cacao mild mosaic virus (CaMMV) using nested PCR and evidence of uneven distribution in leaf tissue. Agronomy 11 (9), 1842. https://doi.org/10.3390/agronomy11091842.

Richardson, S.R., Doucet, A.J., Kopera, H.C., Moldovan, J.B., Garcia-Perez, J.L., Moran, J. V., 2015. The influence of LINE-1 and SINE retrotransposons on mammalian genomes. Microbiol. Spectr. 3 (2), MDNA3-0061–2014. https://doi.org/10.1128/microbiolspec.MDNA3-0061-2014.

Richert-Pöggeler, K.R., Shepherd, R.J., 1997. Petunia vein-clearing virus: a plant pararetrovirus with the core sequences for an integrase function. Virology 236 (1), 137–146. https://doi.org/10.1006/viro.1997.8712.

Robinson, J.T., Thorvaldsdóttir, H., Wenger, A.M., Zehir, A., Mesirov, J.P., 2017. Variant review with the integrative genomics viewer. Cancer Res. 77 (21), e31–e34. https://doi.org/10.1158/0008-5472.CAN-17-0337.

Schmidt, N., Seibt, K.M., Weber, B., Schwarzacher, T., Schmidt, T., Heitkam, T., 2021. Broken, silent, and in hiding: tamed endogenous pararetroviruses escape elimination from the genome of sugar beet (*Beta vulgaris*). Ann. Bot. 128 (3), 281–299. https://doi.org/10.1093/aob/mcab042.

Serfraz, S., Sharma, V., Maumus, F., Aubriot, X., Geering, A.D.W., Teycheney, P.Y., 2021. Insertion of badnaviral DNA in the late blight resistance gene (R1a) of brinjal eggplant (*Solanum melongena*). Front. Plant Sci. 12,. https://doi.org/10.3389/fpls.2021.683681 683681.

Spier Camposano, H., Molin, W.T., Saski, C.A. Sequence characterization of eccDNA content in glyphosate sensitive and resistant Palmer amaranth from geographically distant populations. PLoS One 17 (9), e0260906. https://doi.org/10.1371/journal.pone.0260906.

Tamura, K., Stecher, G., Kumar, S., 2021. MEGA11: molecular evolutionary genetics analysis version 11. Mol. Biol. Evol. 38 (7), 3022–3027. https://doi.org/10.1093/molbev/msab120.

Ullah, I., Daymond, A.J., Hadley, P., End, M.J., Umaharan, P., Dunwell, J.M., 2021. Identification of cacao mild mosaic virus (CaMMV) and cacao yellow vein-banding virus (CYVBV) in cocoa (*Theobroma cacao*) germplasm. Viruses 13 (11), 2152. https://doi.org/10.3390/v13112152.

Wang, K., Tian, H., Wang, L., Wang, L., Tan, Y., Zhang, Z., Sun, K., Yin, M., Wei, Q., Guo, B., Han, J., Zhang, P., Li, H., Liu, Y., Zhao, H., Sun, X., 2021. Deciphering extrachromosomal circular DNA in Arabidopsis. Comput. Struct. Biotechnol. J. 19, 1176–1183. https://doi.org/10.1016/j.csbj.2021.01.043.