Insertion of Badnaviral DNA in the Late Blight Resistance Gene (R1a) of Brinjal Eggplant (Solanum melongena)

Saad Serfraz, Vikas Sharma, Florian Maumus, Xavier Aubriot, Andrew D W Geering, Pierre-Yves Teycheney

frontiers in Plant Science

ORIGINAL RESEARCH
published: 23 July 2021
doi: 10.3389/fpls.2021.683681

Insertion of Badnaviral DNA in the Late Blight Resistance Gene (R1a) of Brinjal Eggplant (*Solanum melongena*)

*Saad Serfraz[1,2,3], Vikas Sharma[4†], Florian Maumus[4], Xavier Aubriot[5], Andrew D. W. Geering[6] and Pierre-Yves Teycheney[1,2]**

[1] CIRAD, UMR AGAP Institut, F-97130, Capesterre-Belle-Eau, France, [2] UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro, Capesterre-Belle-Eau, France, [3] Centre of Agricultural Biochemistry and Biotechnology, University of Agriculture, Faisalabad, Pakistan, [4] URGI, INRAE, Université Paris-Saclay, Versailles, France, [5] Université Paris-Saclay, CNRS, AgroParisTech, Ecologie Systématique Evolution, Orsay, France, [6] Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, Brisbane, QLD, Australia

OPEN ACCESS

Edited by:
Katja R. Richert-Pöggeler, Julius Kühn-Institut, Germany

Reviewed by:
Osamah Alisawi, University of Kufa, Iraq
Richard Kormelink, Wageningen University & Research, Netherlands

***Correspondence:**
Pierre-Yves Teycheney
teycheney@cirad.fr

†Present address:
Vikas Sharma, Institute of Bio- and Geosciences, IBG-1, Biotechnology, Forschungszentrum Jülich GmbH, Jülich, Germany

Specialty section:
This article was submitted to Plant Pathogen Interactions, a section of the journal Frontiers in Plant Science

Received: *21 March 2021*
Accepted: *30 June 2021*
Published: *23 July 2021*

Citation:
Serfraz S, Sharma V, Maumus F, Aubriot X, Geering ADW and Teycheney P-Y (2021) Insertion of Badnaviral DNA in the Late Blight Resistance Gene (R1a) of Brinjal Eggplant (Solanum melongena). Front. Plant Sci. 12:683681. doi: 10.3389/fpls.2021.683681

Endogenous viral elements (EVEs) are widespread in plant genomes. They result from the random integration of viral sequences into host plant genomes by horizontal DNA transfer and have the potential to alter host gene expression. We performed a large-scale search for co-transcripts including caulimovirid and plant sequences in 1,678 plant and 230 algal species and characterized 50 co-transcripts in 45 distinct plant species belonging to lycophytes, ferns, gymnosperms and angiosperms. We found that insertion of badnavirus EVEs along with Ty-1 copia mobile elements occurred into a late blight resistance gene (*R1*) of brinjal eggplant (*Solanum melongena*) and wild relatives in genus *Solanum* and disrupted *R1* orthologs. EVEs of two previously unreported badnaviruses were identified in the genome of *S. melongena*, whereas EVEs from an additional novel badnavirus were identified in the genome of *S. aethiopicum*, the cultivated scarlet eggplant. Insertion of these viruses in the ancestral lineages of the direct wild relatives of the eggplant would have occurred during the last 3 Myr, further supporting the distinctiveness of the group of the eggplant within the giant genus *Solanum*.

Keywords: Endogenous viral elements, badnavirus, *Solanum melongena*, *R1* gene, co-transcripts, phylogeny

INTRODUCTION

Endogenous viral elements (EVEs) originate from viruses with RNA or DNA genomes. They are widespread in eukaryotic genomes (Holmes, 2011; Feschotte and Gilbert, 2012) and may comprise large portions of these genomes: it is estimated that 5–8% of the human genome is composed of endogenous retroviruses (ERVs; Holmes, 2011). Integration is active for viruses encoding an integrase, such as retroviruses, but for others without this protein, such as all plant viruses, integration is either passive through non-homologous end-joining (NHEJ) or mediated by retrotransposons (Geuking et al., 2009; Taylor and Bruenn, 2009; Horie et al., 2010). Some EVEs are infective but the majority are replication-defective because of sequence decay or because host regulation mechanisms have co-evolved to suppress their expression (Mitreiter et al., 1994; Feschotte and Gilbert, 2012).

Endogenous viral elements are important sources of novel genetic information that can ultimately play a significant role in the evolution of the host. For example, syncytin-A, a protein involved in placental development in mammals, is encoded by a gene that was acquired through the endogenization of retroviral *Env* gene (Sha et al., 2000; Dupressoir et al., 2005). ERVs can also modify host gene expression through the contribution of alternate promoters, aberrant splicing, premature termination of transcription or gene disruption (Meyer et al., 2017). In plants, the acquisition of novel functions through the endogenization of viral genes has not yet been formally demonstrated. However, Mushegian and Elena (2015) reported the presence of genes encoding homologs to plant virus movement proteins (MPs) from the 30K superfamily in the genome of almost all euphyllophyte plants and their transcription into mRNAs. They also showed that several of the MP-like coding genes experience positive selection at the codon level, suggesting these genes might be expressed and serve a function in plants (Mushegian and Elena, 2015). Filloux et al. (2015) reported the discovery of endogenous geminivirus like sequences (EGV1) in the genome of *Dioscorea* spp. of the *Enantiophyllum* clade. They provided evidence that functional EGV-expressed replication-associated protein (Rep) were expressed in yams for extended periods following endogenization and that some of them are possibly still functionally expressed in several species of the *Enantiophyllum* clade (Filloux et al., 2015).

In plants, most characterized EVEs originate from viruses with DNA genomes in the families *Caulimoviridae* and *Geminiviridae* (Teycheney and Geering, 2011; Geering et al., 2014; Filloux et al., 2015; Diop et al., 2018; Sharma et al., 2020). Caulimovirid EVEs are widespread in all major plant taxa including club mosses, ferns, gymnosperms and angiosperms (Geering et al., 2014; Diop et al., 2018; Gong and Han, 2018). Only five infective EVEs have been reported in plants. All originate from viruses in the family *Caulimoviridae* (Ndowora et al., 1999; Lockhart et al., 2000; Richert-Pöggeler et al., 2003: Chabannes et al., 2013). It is assumed that non-infectious caulimovirid EVEs must impart some beneficial functions to the plant to be retained over periods as long as millions of years. It has been hypothesized that caulimovirid EVEs confer resistance to cognate exogenous viruses by acting as natural viral transgenes to prime gene silencing either at transcriptional or post-transcriptional levels (Mette et al., 2002; Teycheney and Tepper, 2007; Bertsch et al., 2009). However, copy number often far exceeds that necessary to produce efficient gene silencing and there are very few examples where the corresponding exogenous virus is still extant, making resistance a somewhat null function.

Research on the role of caulimovirid EVEs in normal plant metabolism is still very limited. Endogenous caulimovirid movement protein (MP) genes have been reported in most vascular plants that have had their genomes sequenced (Mushegian and Elena, 2015). There is experimental evidence that downregulation of an endogenous MP in the model plant *Arabidopsis thaliana* can result in a small delay in plant development, and reduces the germination rate, especially at high salt concentrations (Carrasco et al., 2019). In another case study, activation of endogenous petunia vein clearing virus (PVCV) to produce independently replicating virus leads to suppression of post-transcriptional silencing of a gene in the anthocyanin biosynthesis pathway and development of pink blotches in the normally white sections of the bicolored petunia flower (Kuriyama et al., 2020). Impacts of endogenization of viral sequences on genome plasticity, host gene expression or gene disruption through insertional mutagenesis are not yet documented in plants but they are likely to occur considering that such impacts have been reported repeatedly for transposable elements (Bennetzen, 2000; Martin et al., 2009; Zeng and Cheng, 2014).

In this study, we investigated the role of caulimovirid EVEs in plant gene expression through a systematic search for co-transcripts including host and caulimovirid sequences. We provide evidence for such co-transcripts in 45 plant species spanning lycophytes, ferns, gymnosperms and angiosperms. We found that the insertion of badnavirus EVEs resulted in the disruption of orthologs of late blight resistance genes (*R1*) in the cultivated brinjal eggplant (*Solanum melongena*) and its direct relatives. This viral insertion promoted alternative splicing of the co-transcripts and provided an alternate viral promoter. We also report on the identification of three novel badnaviruses for which complete genomes were assembled from EVEs whose insertion was estimated to have occurred during the last 3 million years (Myr).

MATERIALS AND METHODS

Transcriptome Analysis

A search for co-transcripts encompassing viral and plant genomic sequences was done using BLASTX on 939 plant and 13 algal transcriptome shotgun assemblies (TSA) available at NCBI (**Supplementary Table 1**). In addition, 1,468 transcriptome assemblies from 943 plant and 230 algal species from the OneKP database[1] were also included in the transcriptome dataset (**Supplementary Table 1**; Leebens-Mack et al., 2019).

Seventy-two complete genome sequences from viruses in eight genera of family *Caulimoviridae* (*Badnavirus, Caulimovirus, Cavemovirus, Petuvirus, Rosadnavirus, Solendovirus, Soymovirus, Tungrovirus*; Teycheney et al., 2020) and in the tentative genus Florendovirus (Geering et al., 2014) served as queries. tBLASTx-based search was performed using default parameters (except *E* value of 10–5). A library of LTR-retrotransposon sequences was used to filter tBLASTx hits and remove retrotransposon sequences. Transcripts were then annotated using conserved domain search (CDD) database (Lu et al., 2020[2]). An R-based script was designed and used to screen annotated transcripts for the presence of both viral and host domains. All co-transcripts were verified manually using NCBI's CDD browser (Team R Development Core, 2018; Lu et al., 2020). Complete annotated transcripts were aligned and visualized using Easyfig V-2.2.2 ([3]Sullivan et al., 2011). Nucleotide and protein sequence alignments were utilized to extract codon

[1]http://www.onekp.com/

[2]https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi

[3]https://mjsull.github.io/Easyfig/

alignments using PAL2NAL program (Suyama et al., 2006). The Fast unbiased Bayesian approximation (FUBAR) method was implemented to compute dN/dS ratio from codon alignments (Weaver et al., 2018).

Analysis of Alternative Splicing and Quantitative Gene Expression

Unassembled RNA-seq datasets for eggplant and related species (Page et al., 2019) were retrieved from sequence read archive (SRA: **Supplementary Table 1**). Adapter trimming and quality filtering was performed on all datasets. *Solanum melongena* reference genome sequence data (Barchi et al., 2019) was retrieved from the solgenomics server[4]. STAR (v-2.7.3a) was utilized to assemble the transcripts on *S. melongena* scaffold Ch0.05855 (Dobin et al., 2013). Alternative splicing finder (AS Finder; Min, 2013) was used for mapping alternative splicing sites on assembled transcripts. Two kbp sequences upstream each transcript were analyzed using Transcriptional Start Site Plant (TSSPlant), which predicts potential transcription start sites by combining characteristics describing functional motifs and oligonucleotide composition of these sites (Shahmuradov et al., 2017), and Neural Network Promoter Prediction (NNPP), which identifies eukaryotic and prokaryotic promoters in DNA sequences (Reese, 2001). Conserved motifs were identified using the plant *Cis*-Acting Regulatory Elements (CARE) database (Lescot et al., 2002).

Analyses of quantitative gene expression were performed on RNA-seq reads from *S. melongena* (two samples, including one under the name *S. ovigerum*, a primitive domesticate of the eggplant; see Page et al., 2019) as well as on two closely related wild species, *S. incanum* and *S. insanum*. RNA-seq reads were mapped on brinjal eggplant (*S. melongena*) genome scaffold Ch0-5855 using STAR aligner (Dobin et al., 2013). Alignment files and genomic scaffolds served as input for Cufflinks v2.2.1 (Trapnell et al., 2012). The resulting Gene Transfer Format (GTF) files generated from each RNA-seq dataset were merged using Cuffcompare script within the Cufflinks package (Trapnell et al., 2012). Output files were further processed by R-based package cummeRbund v 2.28 (Goff et al., 2012) to quantify the expression of genes and their transcript variants.

Search for Endogenous Caulimovirid Sequences, Reconstruction of Complete Viral Genomes and Phylogenetic Analyses

We searched the genomes of *S. melongena* and the closely related species, *S. aethiopicum* (African scarlet eggplant), for caulimovirid reverse transcriptase (RT) protein sequences. For this, the RT-based protein library built by Diop et al. (2018) from 41 RT domains from distinct viruses representing all the genera in the family *Caulimoviridae* was used as queries for a tBLASTn-based search as described above. Putative integration loci were extended 5 kbp upstream and downstream of integration sites. Each extended locus was filtered for background LTR-retrotransposon sequences as described above. Filtered loci were translated using program translate (version 1.6) and protein sequences of 120 amino acids or more were compared to the initial RT library using BLASTp with default parameters (except *E* value 10^{-5}). The resulting set of endogenous RT sequences was clustered using UCLUST (Edgar, 2010) with identity threshold set at 55%. Loci containing caulimovirid sequences were aligned with sequences from the initial RT library using MUSCLE (version 1.26; Hull, 2009). Maximum likelihood analysis was performed on RT-RNase H1 (RH1) regions using 1000 bootstrap values and phylogenetic trees were built using a 80% identity threshold to delineate viral species (Harrach et al., 2011). Phylogenetic trees were built using IQTree v.2.0, choosing the best-fit model of evolution (command -m GTR + F + I + G4) and 1000 bootstrap replicates (Nguyen et al., 2015). Dioscorea nummularia-associated virus (DNUaV) was used as an outgroup. Phylogenetic trees were also reconstructed using Bayesian inference methods. The best-fitting nucleotide substitution model for Bayesian inference was selected based on the Akaike information criterion computed by MrModeltest v.2.3 (Nylander et al., 2004); the GTR + G + I model was chosen and used in MrBayes 3.2.6 (Huelsenbeck and Ronquist, 2001). MrBayes analysis constituted two independent parallel runs of four Markov chains each, implemented for 5 million generations and sampled every 500 generations. Adequate mixing of the Markov chains and convergence of the two runs were confirmed with Tracer v1.6 (Rambaut et al., 2014). After removing a 10% burnin, the remaining trees were used to generate the 50% Bayesian majority-rule consensus tree. The reconstruction of complete badnavirus genomes from EVEs was carried out using previously described methods (Geering et al., 2014). The recombination detection program RDP v 4 (Martin et al., 2015) was used with default settings to detect recombination signals. Viral sequences were initially aligned by MUSCLE (version 1.26; Hull, 2009) and manually edited for recombination analysis.

Synteny and Phylogenetic Analyses

R1 ortholog loci identified in *S. aethiopicum*, *S. demissum*, *S. melongena*, *S. tuberosum*, and *S. verrucosum* were extended manually up to 15 kbp upstream and downstream. Reciprocal BLASTn analyses were performed on these extended loci. Synteny analyses were performed using Easyfig V-2.2.2 (Sullivan et al., 2011), with conserved genes located downstream *R1* genes as chromosomal anchors.

A library of NB-ARC domains was constructed from reference plant resistance (*R*) genes of the PRG database (Sanseverino et al., 2009) and used to search NB-ARC homologs in chromosomal contigs from *S. aethiopicum*, *S. commersonii*, *S. melongena*, *S. tuberosum*, and *S. verrucosum* using tBLASTn. Output sequences were translated using program translate (version 1.9) and aligned using MUSCLE. Resulting MSA files were used to build phylogenetic trees as described above except best-fit model of evolution (command – m JTT + F + R4).

A library of RT domains from transposable elements (TEs) was built by clustering reference TEs from repeat explorer database (Novák et al., 2013) at 50% identity threshold. This TE library served as queries to search each contig from *Solanum* species

[4]ftp://ftp.solgenomics.net/genomes/

using tBLASTn. Phylogenetic comparisons of output TEs were performed as described above except best-fit model of evolution (command – m VT + F + G4). LTRHarvest was also utilized for LTR prediction and annotation (Ellinghaus et al., 2008).

Plant Material, PCR Amplification and Sequencing

Seeds from 26 eggplant accessions and their wild relatives were provided by INRAE-Centre for Vegetable Germplasm ([5]**Supplementary Table 2**). Seed-lines were selected to include nine of the 13 currently recognized species of the Eggplant clade, as delimited by Knapp et al. (2013) and Aubriot et al. (2018), including the wild progenitor of the eggplant, *S. insanum*. Seven species from the Anguivi grade (viz. a large and unresolved Afro-Asian lineage that includes the Eggplant clade and *S. aethiopicum*, the cultivated scarlet eggplant; see Aubriot et al., 2016) were also sampled (Knapp et al., 2013; Aubriot et al., 2016). Seeds were soaked for 24 h at room temperature in Petri dishes containing 500 ppm of gibberellic acid-3 (GA3), transferred to pots containing potting mix and let germinate at room temperature. Leaf samples were collected from seedlings at the five-leaf stage and stored at −80°C until further use. Total DNA was extracted using FastDNA kit (MP Bio, Irvine, CA, United States) and used to search for Caulimovirid EVE insertion loci by PCR using primer pair InL_F2 (CAAACAATACGGTACAACTC)/InL_R2 (CAACAGCATCGATGAATTC; **Supplementary Table 3**). These primers were designed based on the alignment of the flanking regions of the viral insertion locus of *S. melongena* (V*R1*) and the corresponding sequences in the genome of *S. aethiopicum*, which has no viral insertion. PCRs were done using Phusion DNA polymerase (NEB, Ipswich, MA, United States). PCR mixes with a final volume of 25 μl contained 2.5 μl of 10 × GC buffer, 0.2 μl of 25 mM $MgCl_2$, 1.6 μl of 2.5 mM dNTPs, 1 U of 5,000 U/ml Phusion DNA polymerase and 5 μl of each primer at a concentration of 5 μM. Thermocycling conditions were an initial denaturation of 5 mn at 98°C, 35 cycles of 10 s at 98°C, 15 s at 63°C and 3 mn at 72°C followed by a final extension of 5 mn at 72°C. Amplification products were purified using QIAquick gel extraction kit (QIAGEN, Toronto, ON, Canada) and sequenced by Genewiz (Leipzig, Germany). In order to amplify upstream and downstream border regions, an additional set of primers (R1U: CTCGATTATGCCAAATCCTC and R1D: GCTTGAAGTATGCGAGAAG) was designed within the viral insertion of *S. melongena* (**Supplementary Table 3**). The left and right borders of the V*R1* insertion locus were amplified by PCR using the primer pairs InL_F2/R1U and lnLR2/R1D, respectively. PCR mixes with a final volume of 25 μl contained 2.5 μl of 10 × reaction buffer (10 mM Tris–HCl, 50 mM KCl, 1.5 mM $MgCl_2$, pH 8.3), 0.75 μl of 25 mM $MgCl_2$, 1 μl of 2.5 mM dNTPs, 1 U of 5,000 U/ml *Taq* DNA polymerase (NEB, Ipswich, MA, United States) and 2.5 μl of each primer at a concentration of 2 μM. PCR conditions were an initial denaturation of 5 mn at 94°C and 35 cycles of 15 s 94°C for, 30 s at 54°C and 2 mn at 72° followed by a final extension of 5 mn at 72°C. Amplification products were cloned using pGEM®-T Easy Vector Systems (Promega, Charbonnières-les-Bains, France) and sequenced by Genewiz (Leipzig, Germany).

[5]https://www6.paca.inrae.fr/gafl_eng/Vegetable-Germplasm-Centre

RESULTS

Search for Co-transcripts

We used a tBLASTx-based pipeline to search the transcriptomes of 1,678 plant and 230 algal species for potential co-transcripts including host and caulimovirid sequences. Fifty distinct co-transcripts were identified from the transcriptomes of 45 plant species representing lycophytes, ferns, gymnosperms, and angiosperms (**Table 1**). These co-transcripts include sequences encoding partial or complete caulimovirid MP, reverse-transcriptase/ribonuclease H (RT-RH) or aspartic protease (AP). Co-transcripts including caulimovirid MPs and host chaperone sequences were identified in several distinct species (**Table 1**). A co-transcript encoding a soymovirus polyprotein with coat protein (CP), AP, RT, RH1, and MP domains and a long chain acyl-CoA synthetase 1 domain was identified in alfalfa (*Medicago sativa*; **Table 1** and **Figure 1**).

Co-transcripts including caulimovirid RT-RH sequences and *R1* late blight resistance gene sequences were identified in brinjal eggplant (*S. melongena*) and in closely related species from the Eggplant clade (*S. incanum*; **Table 1** and **Figure 1**). Similar co-transcripts were identified in the wild progenitor of the eggplant, *S. insanum*, from transcriptomes of the SRA database (Page et al., 2019). We selected these co-transcripts for further investigation because genomic and transcriptomic resources are available for their plant host species and because identical co-transcripts were identified in the transcriptomes of these three closely related *Solanum* species (**Figure 1**).

Co-transcripts identified in *S. melongena*, *S. incanum* and *S. insanum* contained caulimovirid reverse transcriptase (RT) and RNAseH1 (RH1) domain sequences, and *R1* NB-ARC and LRR domain sequences (**Figure 2**). They ranged in size between 4 and 8 kb. Insertion of viral sequences resulted in multiple disruptions in *R1*, making translation into functional proteins unlikely. Mapping co-transcripts on the reference genome of *S. melongena* revealed a single insertion locus for this species, in which a 2.4 kbp sequence in the *R1* NB-ARC domain was replaced by a 5.8 kbp caulimovirid sequence (**Figure 2**). We named this locus virus *R1* (V*R1*). It is located at positions 432,102–443,945 bp on contig 05585 of S. *melongena* chromosome 0.

Analysis of VR1 Co-transcripts

RNA-seq datasets from one *S. incanum*, two *S. insanum* and ten *S. melongena* accessions were used for analyzing V*R1* co-transcripts. Transcripts of different lengths (4–8 kb) were identified for the V*R1* locus of these three species, suggesting that alternative splicing occurs (**Figure 2**). Further evidence for alternative splicing was gathered using Splice Aware Aligner (STAR) and AS Finder-based pipeline. Five co-transcript variants (isoforms) were identified in *S. melongena*, five in *S. incanum* and two in *S. insanum* (**Figure 2**). The structure of these variants suggests that intron skipping and intron retention occurs in *S. incanum* and *S. melongena* and shows that *R1* highly conserved

TABLE 1 | List of co-transcripts containing caulimovirid and plant sequences identified in this work.

Source	ID	Species	Family	Group	Number of co-transcripts	Host domain/function	Virus domain
OneKp	EZZT	*Passiflora edulis*	*Passifloraceae*	Rosids	1	V-type ATP synthase subunit I	AP
TSA, NCBI	GGXZ02	*Platanus x hispanica*	*Platanaceae*	Basal Eudicots	2	Archaeal histone H3/H4	AP
OneKp	BDJQ	*Zingiber officinale*	*Zingiberaceae*	Commelinids	1	SH3 domain protein	AP,RT
OneKp	HDSY	*Aerva persica*	*Amaranthaceae*	Core Eudicots	1	Laminin G domain	CP,AP,RT,RH
OneKp	TOKV	*Aphanopetalum resinosum*	*Aphanopetalaceae*	Core Eudicots	1	glycosyltransferase family 1 and related proteins with GTB topology	MP
OneKp	RKGT	*Eschscholzia californica*	*Papaveraceae*	Basal Eudicots	4	GTP-binding nuclear protein Ran	MP
OneKp	GTUO	*Huperzia selago*	*Huperziaceae*	Lycophytes	1	light-harvesting complex chlorophyll-a/b protein of photosystem I (Lhca)	MP
OneKp	XLIQ	*Iodes vitiginea*	*Icacinales*	Asterids	1	light-harvesting complex chlorophyll-a/b protein of photosystem I (Lhca)	MP
OneKp	TJES	*Spergularia media*	*Caryophyllaceae*	Core Eudicots	1	PPR repeat family	MP
OneKp	XSSD	*Amaranthus hybridus*	*Amaranthaceae*	Core Eudicots	2	molybdopterin-synthase adenylyltransferase MoeB	MP
OneKp	YUOM	*Toxicodendron radicans*	*Anacardiaceae*	Rosids	1	Polysaccharide Lyase Family 6	MP
OneKp	IPPG	*Heliotropium filiforme*	*Boraginaceae*	Asterids	3	Aldolase/RraA	MP
OneKp	SCEB	*Podocarpus coriaceus*	*Podocarpaceae*	Conifers	1	OPT oligopeptide transporter protein	MP
OneKp	PVGM	*Oncotheca balansae*	*Oncothecaceae*	Asterids	1	Ubiquitin-conjugating enzyme E2, catalytic (UBCc) domain.	MP
TSA, NCBI	GFVC01	*Jatropha curcas*	*Euphorbiaceae*	Rosids	1	Heat shock protein 70 (HSP70)	MP
OneKp	PUCW	*Agastache rugosa*	*Lamiaceae*	Asterids	2	Molecular chaperone IbpA, HSP20 family	MP
OneKp	WHNV	*Clinopodium serpyllifolium*	*Lamiaceae*	Asterids	2	Molecular chaperone IbpA, HSP20 family	MP
TSA, NCBI	GFRR01	*Ocimum tenuiflorum*	*Lamiaceae*	Asterids	1	Molecular chaperone IbpA, HSP20 family	MP
TSA, NCBI	GFBP01	*Juncus effusus*	*Juncaceae*	Commelinids	1	Molecular chaperone IbpA, HSP20 family	MP
TSA, NCBI	GALV01	*Gossypium hirsutum*	*Malvaceae*	Rosids	2	Molecular chaperone IbpA, HSP20 family	MP
TSA, NCBI	IADW01	*Orobanche minor*	*Orobanchaceae*	Asterids	5	Molecular chaperone IbpA, HSP20 family	MP
TSA, NCBI	GDKT01	*Vigna unguiculata*	*Fabaceae*	Rosids	1	LRR kinase	MP
TSA, NCBI	GDIZ01	*Silene conica*	*Caryophyllaceae*	Core Eudicots	1	Acyl transferase	MP
TSA, NCBI	GDJH01	*Silene conica*	*Caryophyllaceae*	Core Eudicots	1	acyl transferase	MP
TSA, NCBI	GCZN01	*Abies pinsapo*	*Pinaceae*	Conifers	1	Phage-related minor tail protein	MP
TSA, NCBI	GFLL01	*Pinus albicaulis*	*Pinaceae*	Conifers	2	enoyl_reductase_like	MP
TSA, NCBI	GGKA01	*Medicago sativa*	*Fabaceae*	Rosids	3	Long-chain acyl-CoA synthetase	MP,RT,RH,CP
OneKp	TPEM	*Platyspermation crassifolium*	*Escalloniaceae*	Asterids	1	Major intrinsic protein (MIP) superfamily.	RH
OneKp	WPHN	*Idiospermum australiense*	*Calycanthaceae*	Magnoliids	1	Peroxiredoxin (PRX) family,	RH
OneKp	YYPE	*Austrocedrus chilensis*	*Cupressaceae*	Conifers	1	Seed maturation protein.	RH
TSA, NCBI	GDIY01	*Silene conica*	*Caryophyllaceae*	Core Eudicots	3	RING-finger-containing ubiquitin ligase	RH
TSA, NCBI	GDQW01	*Panax ginseng*	*Apiineae*	Asterids	6	Cyclic nucleotide-binding domain.	RH
OneKp	LWCK	*Lycium barbarum*	*Solanaceae*	Asterids	1	molecular chaperone DnaK	RT

(Continued)

TABLE 1 | Continued

Source	ID	Species	Family	Group	Number of co-transcripts	Host domain/function	Virus domain
OneKp	JOPH	*Carapichea ipecacuanha*	*Rubiaceae*	Asterids	1	Photosystem I psaA/psaB protein	RT
OneKp	ZSGF	*Carapichea ipecacuanha*	*Rubiaceae*	Asterids	1	Photosystem I psaA/psaB protein	RT
OneKp	UFJN	*Diplazium wichurae*	*Athyriaceae*	Ferns	1	C-terminal domain of rhamnogalacturonan lyase	RT
OneKp	TJQY	*Kerria japonica*	*Rosaceae*	Rosids	1	chromosome segregation protein SMC	RT
OneKp	SLYR	*Cladrastis kentukea*	*Fabaceae*	Rosids	1	Sel1-like repeats	RT
OneKp	LQJY	*Solanum virginianum*	*Solanaceae*	Asterids	1	Tetratricopeptide repeat	RT
OneKp	QSNJ	*Taiwania cryptomerioides*	*Cupressaceae*	Conifers	1	core domain of the SPFH (stomatin, prohibitin, flotillin, and HflK/C) superfamily	RT
TSA, NCBI	GFZE01	*Catharanthus roseus*	*Apocynaceae*	Asterids	2	LRR Kinase protein	RT
TSA, NCBI	GBHJ01	*Withania somnifera*	*Solanaceae*	Asterids	3	GTPase domain	RT
TSA, NCBI	GGDV01	*Croton tiglium*	*Euphorbiaceae*	Rosids	5	leucine-rich repeat receptor-like protein kinase	RT
OneKp	KXSK	*Agave tequilana*	*Agavaceae*	Monocots	1	Trimeric dUTP diphosphatases	RT,RH
OneKp	UQCB	*Portulaca molokiniensis*	*Portulacaceae*	Core Eudicots	1	Cupredoxin superfamily	RT,RH
OneKp	UQCB	*Portulaca molokiniensis*	*Portulacaceae*	Core Eudicots	1	Largest subunit of RNA polymerase (RNAP), C-terminal domain	RT, RH
OneKp	KEGA	*Glycine soja*	*Fabaceae*	Rosids	1	*P. syringae* resistance	RT,RH
TSA, NCBI	GAYS01	*Solanum incanum*	*Solanaceae*	Asterids	2	Late blight resistance gene R1	RT,RH
TSA, NCBI	GAYR01	*Solanum melongena*	*Solanaceae*	Asterids	5	Late blight resistance gene R1	RT,RH
TSA, NCBI	GBEF01	*Solanum melongena*	*Solanaceae*	Asterids	2	Late blight resistance gene R1	RT,RH

intron 1 (Ballvora et al., 2002) was expressed in *S. melongena* and *S. incanum* but not in *S. insanum* (**Figure 2**). Transcripts expressed in all three species retained the second intron located in the *R1* LRR domain (**Figure 2**). Short (4 kb) and long (8 kb) co-transcripts were mapped to the same insertion loci, suggesting the existence of alternative viral promoters in the V*R1* locus. Therefore, a search was done for alternative promoters in the V*R1* locus of *S. aethiopicum* and *S. melongena*. Two putative promoters (P1 and P2) were identified, potentially driving the synthesis of long and short transcripts, respectively (**Figure 2**). The P1 promoter is located upstream of the transcription start site (TSS) of the long transcripts. It displays 91.7% sequence identity with the closest homolog of the *R1* promoter of *S. aethiopicum* and *S. melongena*. The P2 promoter was identified upstream of the TSS of short transcripts located within the viral insertion and is therefore of viral origin. The TSS and NDPP programs predicted a promoter with linear discriminant factor scores of 1.977 (threshold = 1.52), and 0.97 (cutoff = 0.8), respectively. PlantCARE identified CAAT box motifs and a TATA-box located at 699 and 639 bp upstream of the TSS, respectively.

The transcription level of each co-transcript was assessed in one *S. incanum*, two *S. insanum*, and ten *S. melongena* accessions using the RNA-seq pipeline Tuxedo (Trapnell et al., 2012). Reads were mapped on *S. melongena* scaffold Ch0.5855 (Barchi et al., 2019). Transcription levels were very similar for all analyzed cultivars, except for *S. melongena* accessions Mel-4 and Mel-5, which displayed higher expression levels (**Supplementary Figure 1**). They were also similar to the transcription level (60,127.13 FPKM) of a constitutively expressed gene (ORC-6). These data suggest that co-transcripts were expressed constitutively in all analyzed *S. melongena*, *S. incanum* and *S. insanum* accessions.

Classification of the EVE at the VR1 Locus

Endogenous viral elements at the V*R1* locus of *S. incanum*, *S. insanum*, and *S. melongena* had 100% nucleotide identity in the RT/RH1 coding regions, suggesting they derive from the same ancestral virus, for which the name brinjal badnavirus A (BBVA) is proposed (**Supplementary Table 3**). When a BLASTX search of the NCBI non-redundant protein database was done, matches were obtained to a range of badnaviruses, with the highest scoring match being to blackberry virus F (Sequence ID: YP_009229919.1). A near complete consensus genome of BBVA was assembled from the EVEs located on contig Ch0_17757 of *S. melongena*. The genome organization of BBVA is similar to that of badnaviruses[6], including an open reading frame (ORF) 3 with putative MP, CP, AP, RT, and RH1 coding regions.

[6]https://talk.ictvonline.org/ictv-reports/ictv_online_report/reverse-transcribing-dna-and-rna-viruses/w/caulimoviridae/1361/genus-badnavirus

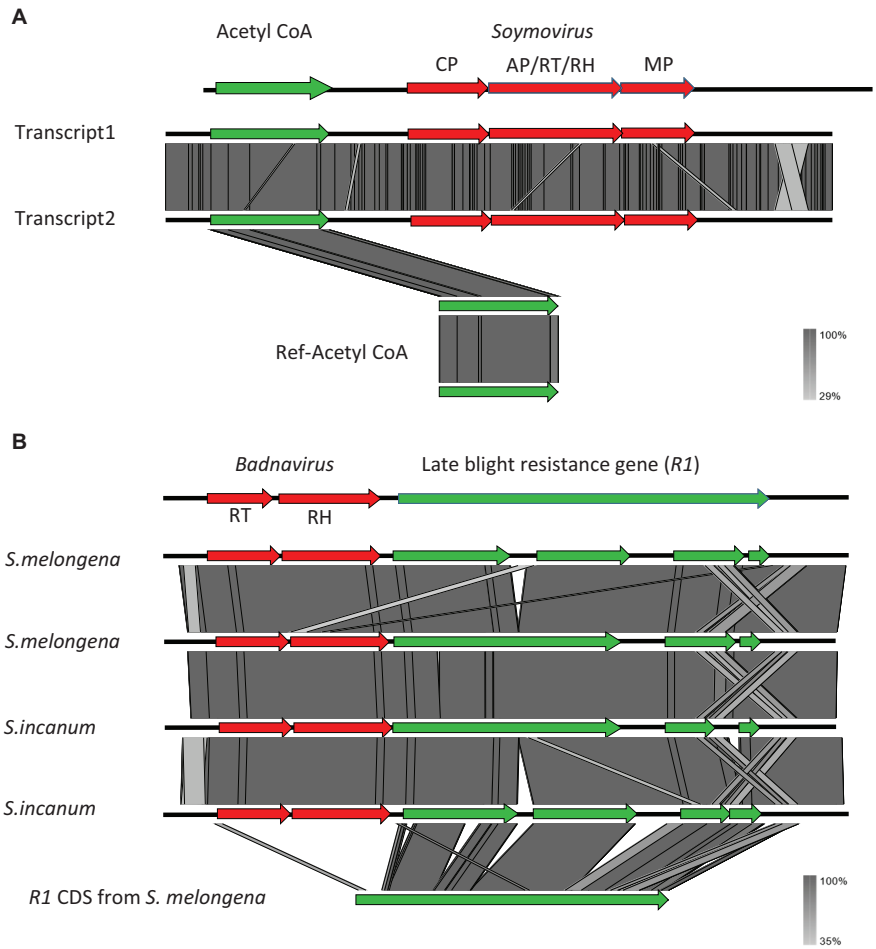

FIGURE 1 | Structure of the co-transcripts identified in alfalfa (*Medicago sativa*) **(A)** and eggplant (*Solanum melongena*) and *Solanum incanum* **(B)**. Viral sequences are shown in red and host sequences are shown in green.

The reconstructed sequence of BBVA was used to search for related sequences in the recently sequenced nuclear genomes of *S. melongena* and its African relative *S. aethiopicum* (Barchi et al., 2019; Song et al., 2019). An additional badnaviral EVE was identified in *S. melongena* chromosome 1 (91008101–91010200) which only had 76% nt identity to the aforementioned sequences of BBVA in the RT/RNase H region and hence could be regarded as representing a different ancestral badnavirus, for which we propose the name brinjal badnavirus B (BBVB). Yet another distinct badnaviral sequence was identified in *S. aethiopicum*, for which the name gilo badnavirus (GBV) is proposed. This name is derived from gilo, the vernacular name of *S. aethiopicum*.

To confirm classification, phylogenetic analyses were done using all EVEs that were found. The topologies of the trees obtained using maximum likelihood (**Figure 3**) and Bayesian methods (**Supplementary Figure 2**) were similar, with the EVEs included within a larger clade of eudicot-infecting badnaviruses originating mainly from the Old World. No evidence of recombination between BBVA, BBVB, GBV and other members of genus *Badnavirus* could be found following the analysis of badnavirus genome sequence alignments using the recombination detection program RDP4 (Martin et al., 2015).

Endogenization of BBVA

Searches for BBVA EVEs were extended to 11 additional *Solanum* species from the tomato clade (*S. chilense*, *S. galapagense, S. lycopersicum, S. pennelli, S. peruvianum*, and *S. pimpinellifolium*), the potato clade (*S. demissum, S. tubersoum*, and *S. verrucosum*) and the Eggplant clade (*S. incanum*, *S. insanum*). For this, BlastN analyses were performed on complete genome sequences available for species in the tomato and potato clades using the assembled BBVA genome sequence as bait. Similar analyses were performed on transcriptomic datasets available for *S. incanum* and *S. insanum*, for which no genome sequence data is available. All screened *Solanum* species were devoid of BBVA insertions except *S. incanum* and *S. insanum* as mentioned above.

Solanum melongena genomic sequences surrounding BBVA sequences at the V*R1* locus are flanked by an *R1* ortholog, sterol-C5(6)-desaturase (ERG3) (PFAM ID: PF01222) and ORC-6 (PFAM ID: PF05460) (**Figure 4**). Syntenic relationships were observed for BBVA, *R1* and ERG3 sequences between members

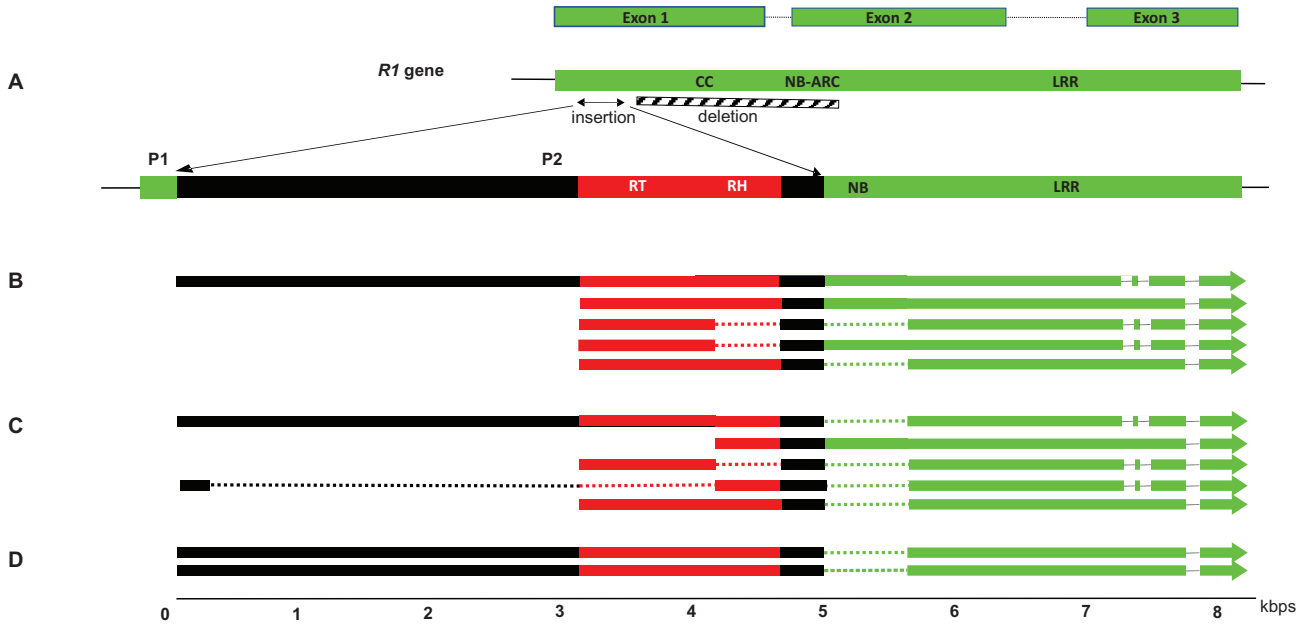

FIGURE 2 | Structure of the VR1 locus of *S. melongena* **(A)** and co-transcripts identified in *S. incanum* **(B)**, *S. melongena* **(C)** and *S. insanum* **(D)**. Viral and host sequences are shown in red and green, respectively. Spliced sequences in the co-transcripts are represented by dotted lines and the deletion at *S. melongena* VR1 locus is represented as a hatched box. The position of alternate promoters P1 and P2 are shown.

of the Eggplant clade. Synteny was also observed for *R1*, ERG3 and ORC-6 between the genome of *S. melongena* and those of *S. aethiopicum*, *S. demissum*, *S. tuberosum*, and *S. verrucosum*, which are devoid of endogenous BBVA sequences. The V*R1* loci is very conserved between *S. incanum*, *S. insanum*, and *S. melongena* with nucleotide sequence identities of 98–99%.

PCR-based screenings were performed on *Solanum* species directly related to *S. melongena* (i.e., members of the Eggplant clade and Anguivi grade for which genome sequences are not available) using primer pair InlF2/InlR2, which allow the amplification of the entire V*R1* locus (**Supplementary Table 4**). Twenty-six accessions representing 17 species were screened (**Supplementary Table 2**). Amplification products of the expected size (6,866 bp) were obtained for 18 accessions representing seven species from the Eggplant clade (*S. incanum*, *S. insanum*, *S. linnaeanum*, *S. melongena*, *S. rigidum*, and *S. umtuma*) indicating the presence of endogenous BBVA sequences in the genomes of these species (**Supplementary Figure 3B**). Smaller amplicons (∼6,500 bp) were obtained for *S. campylacanthum* and *S. lichtensteinii*, whereas two amplicons of 4 and 1.5 kbp were obtained for *S. cerasiferum*. Attempts to clone these PCR products were unsuccessful due to their large sizes. Amplification products of 1.5 kbp were also obtained for species from the Anguivi grade (*S. aethiopicum*, *S. anguivi*, *S. dasyphyllum*, *S. macrocarpon, S. richardii*, *S. trilobatum*, and *S. violaceum* [see Aubriot et al. (2018) for details on these species] using the same primer pair (**Supplementary Figure 3C**). These amplification products were cloned and sequenced. Sequence analyses showed that they were devoid of BBVA sequences. These sequences were used in the synteny analyses illustrated in **Figure 5**.

These results suggest that endogenization of BBVA sequences could be a derived trait for the Eggplant clade and that a hypothetical ancestral lineage of the Eggplant might already have an endogenized BBVA sequence: BBVA insertion would have followed the divergence of the lineage formed by the eggplant and its closely related wild species (the Eggplant clade) from the other *Solanum* lineages (including the closely related Anguivi grade that accounts for *S. aethiopicum*).

Additional PCRs were performed using primer pairs InlF2/R1U and InrF2/R1D (**Supplementary Table 4** and **Figure 5**) to amplify the left and right borders of the V*R1* locus, respectively, from the six species for which the complete VR1 locus was amplified (*S. incanum*, *S. insanum*, *S. linnaeanum*, *S. melongena*, *S. rigidum*, and *S. umtuma*). Amplicons corresponding to the right borders could not be obtained, despite several attempts and the design and use of alternative primers to primer R1D in the viral part of the right border. These alternative primers and primer R1D target the highly conserved NB-ARC domain of *R1* genes and other NB-LRR genes that are scattered across the genome of *S. melongena* (Barchi et al., 2019). This might explain why specific PCR products could not be obtained, thus preventing the sequencing of the right border. In contrast, amplicons of the expected size (∼1,545 bp) corresponding to the left border were obtained for all six species (**Supplementary Figure 3D**). These amplicons, which include the junction between the *R1* gene and BBVA sequences, were cloned and sequenced. Sequence comparisons

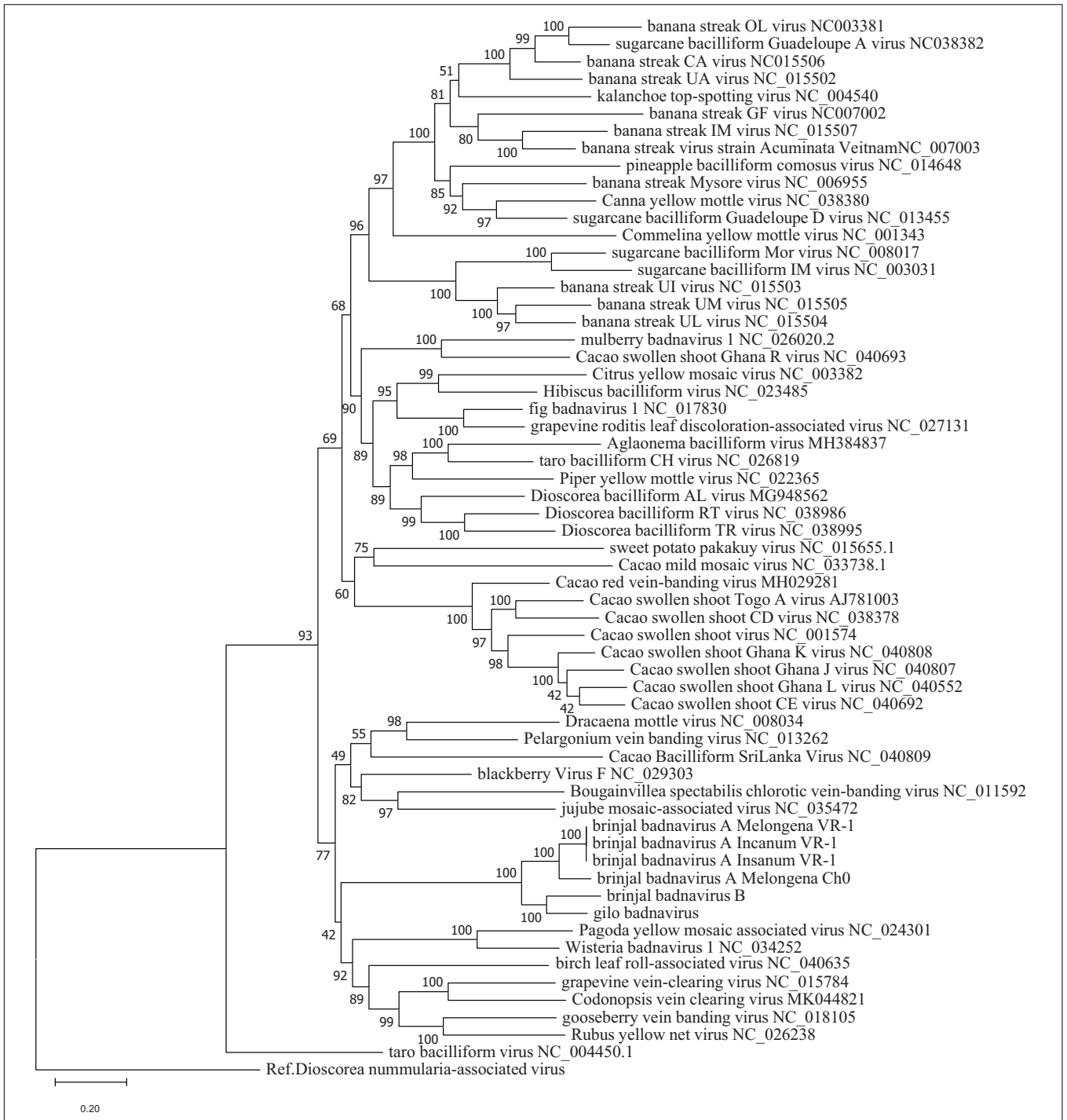

FIGURE 3 | Phylogenetic tree showing placement of BBVA, BBVB, and GBV in genus *Badnavirus* using ML method. Analysis with 1000 bootstrap sets were performed on nucleotide sequences of the RT-RNase H domains corresponding to positions 5,599–6,882 in the genome of Commelina yellow mosaic virus (NC_001343), the type member of genus *Badnavirus*. Sequences of the proposed new species are shown in light red box.

showed that they display 84.4–99.1% nucleotide identity to each other (**Supplementary Data File 1**).

Analysis of the synteny between genomic regions corresponding to the left border of *S. melongena* V*R1* locus and comprising the *R1*/BBVA junction was carried out for all *Solanum* species for which this region could be amplified, cloned and sequenced (**Figure 5**). It showed that synteny between 8 species belonging to three of the four subclades within the Eggplant clade (widespread clade: *S. campylacanthum*, *S. cerasiferum* and *S. incanum*; Southern African clade: *S. lichtensteinii*,

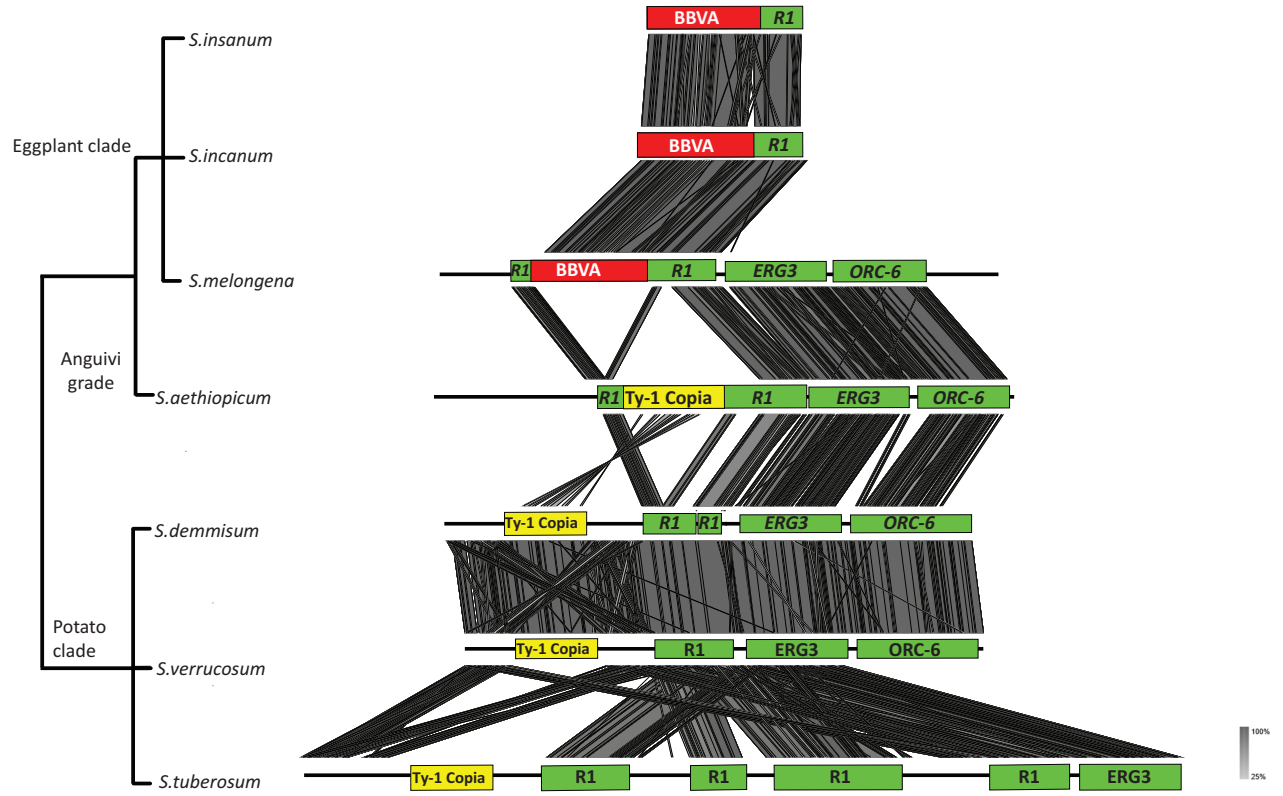

FIGURE 4 | Syntenic relationships between S. *melongena VR1* locus and homologous genomic regions of related *Solanum* species. Viral sequences are shown in red and host sequences are shown in green. Ty-1 copia transposable elements are shown in yellow boxes. ERG3: sterol-C5(6)-desaturase; ORC-6: origin recognition complex 6.

S. linnaeanum, and *S. umtuma;* eggplant and wild relative: *S. insanum* and *S. melongena*) and five species from the Anguivi grade (*S. aethiopicum*, *S. anguivi*, *S. dasyphyllum*, *S. macrocarpon*, *S. richardii*, *S. trilobatum*, and *S. violaceum*) is coherent with the phylogenetic framework proposed by Aubriot et al. (2018) for the Eggplant clade.

DISCUSSION

The contribution of EVEs to the biology of their hosts through structural and/or functional modifications of their genomes is well documented in mammals (Feschotte and Gilbert, 2012) but it is more limited in plants. However, the widespread presence of caulimovirid EVEs in plant genomes (Geering et al., 2014; Diop et al., 2018) and their conservation throughout plant evolution raise questions about their contribution to plant biology. In this study, we performed a comprehensive search for co-transcripts including plant and caulimovirid sequences in publicly available databases. We identified 50 such co-transcripts in 45 plant species belonging to 31 distinct families of vascular plants, providing evidence that fused ORFs including caulimovirid sequences are present in plant genomes.

Three of these co-transcripts were characterized in brinjal eggplant (*Solanum melongena*), and its close relatives *S. insanum* and *S. incanum*, providing the first evidence of badnavirus infections in the *Solanaceae* since no extant badnavirus has yet been reported for this economically important family. According to our analyses, BBVA inserted into gene *R1*, a NBS-LRR gene involved in resistance against *Phytophtora infestans*. *R1* belongs to the multigenic leucine zipper/NBS/LRR class of plant resistance genes and confers resistance against *P. infestans* in potato (Ballvora et al., 2002; Kuang et al., 2005) and other *Solanum* species including *S. melongena*. *R1* is used for breeding late blight-resistant crop cultivars (Faino et al., 2010), although *R1*-mediated resistance introgressed through breeding can be rapidly overcome by new strains of *P. infestans* (Fry and Goodwin, 1997). *R1* genes typically encode a 1,293 amino-acid residue polypeptide containing a coiled coil (CC) domain, a nucleotide-binding ARC domain (NB-ARC) and a leucine rich repeats (LRR) domain (**Figure 2**). Given the divergence dates and biogeographical reconstructions obtained by Aubriot et al. (2018), we suggest that BBVA endogenization has probably occurred once, in northern Africa or the Middle-East region during the last ∼ 3 Myr. *P. infestans* is considered to have originated from Mexico (Grünwald and Flier, 2005; Goss et al., 2014) and therefore there would not have been encounters between this pathogen and eggplant until relatively recently, certainly not before the discovery of the New World by Christopher Colombus in 1492. Hence, at the time of integration

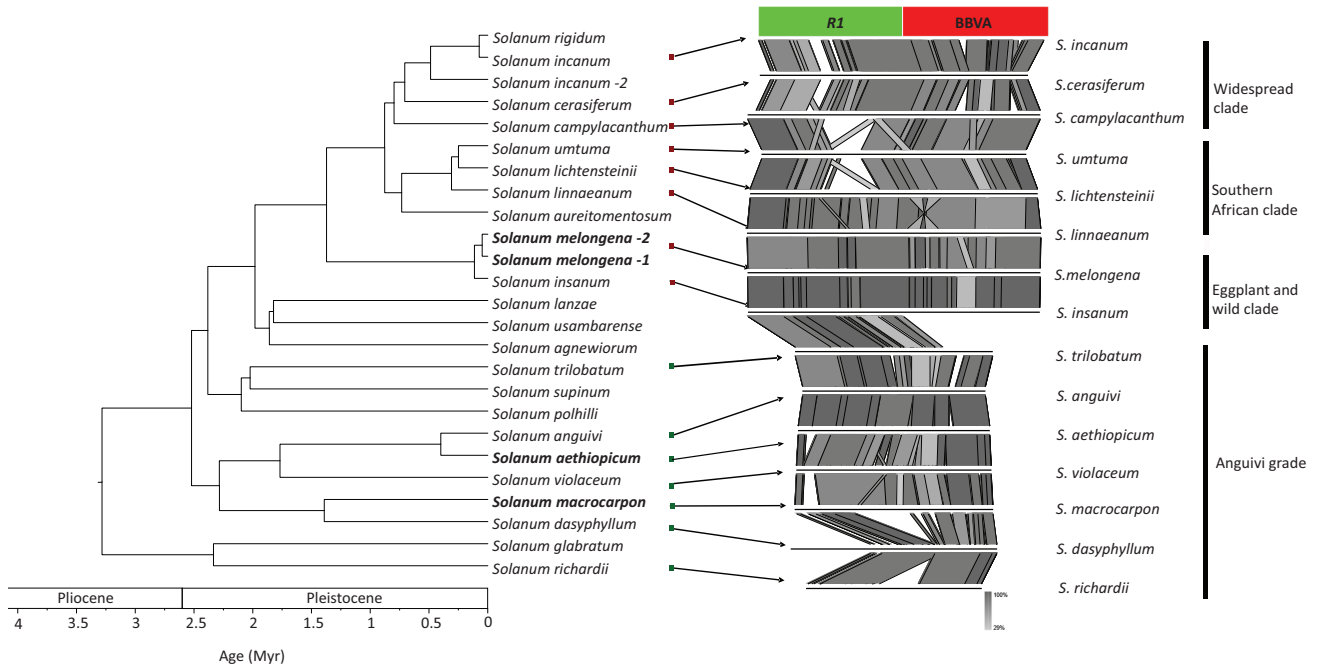

FIGURE 5 | Phylogenetic relationships of *S. melongena* and related *Solanum* species along with syntenic relationships of the left border of the V*R1* locus. Left part of the figure reproduced from Aubriot et al., 2018).

of BBVA, the plant population would not have been under selection pressure from *P. infestans* and gene *R1* may have been redundant in function. Endogenization of BBVA may have contributed to diversification of this gene family, and provided some positive benefit to the plant, particularly as this endogenous viral element has been retained through several plant speciation events.

Although splicing was reported before in the expression of rice tungro bacilliform virus (RTBV), a member of family *Caulimoviridae* (Fütterer et al., 1994), the impact of EVEs on alternate splicing is not yet documented in plants. Here, we provide evidence that the insertion of BBVA has disrupted *R1* and caused intron splicing. Transcriptomic analyses showed that several spliced transcripts of *S. melongena* and *S. incanum* display exon skipping and intron retention. Alternative splicing is known to be involved in the rearrangement of domains with preexisting functions into new protein composite architectures through exon shuffling, resulting in the acquisition of important new functions in eukaryotes hosts such as host-transposase fusion (HTF) genes (Cosby et al., 2021). Our work provides evidence that caulimovirid EVEs can promote alternative splicing in plants and have the potential to help forming fusion proteins with new functionalities. Premature adenylation was not observed in any co-transcript as all share a common 3′ end sequence. In contrast, two different 5′ ends were identified, suggesting the existence of alternative promoters. *In silico* promoter analysis suggested that a putative promoter (P1) from the *R1* gene might be driving the expression of long transcripts (8 kb) whereas a putative viral promoter (P2) identified within the BBVA insertion might be driving the expression of shorter transcripts (4–5 kb). Conserved caulimovirid *cis*-elements (TATA box and CAAT box) were identified in P2. Although co-option of viral promoters from EVEs was not reported before in plants, that of promoters from ERVs by animal genomes is known to contribute to the regulation of the expression of mammalian genes and to the transcription of non-coding genes (ncRNAs). Promoters acquired from ERV LTR regions also function as alternative and tissue-specific promoters (Faulkner et al., 2009; Beyer et al., 2011), whereas some retrotransposon internal coding sequences can serve as promoters.

Fast Unconstrained Bayesian AppRoximation (FUBAR, Murrell et al., 2013) analyses were performed on the RTs from BBVA, BBVB and GBV. Posterior probabilities $\geqq$ 0.9, were registered for 13/40 sites, providing evidence of purifying selection (**Supplementary Table 5**). These results suggest that badnaviral RT sequences are conserved in brinjal eggplant and its wild relatives and could potentially be translated into functional proteins.

As evolution of viral lineages are spatiotemporally coupled with divergence and spread of host population, they can bring novel types of data to the traditional phylogenetic context. Hence, comparisons of shared BBVA insertion confirm *S. insanum* as the closest relatives of the eggplant and corroborate the current delimitation of the Eggplant clade. Current data suggest that BBVA EVEs in the genome of the Eggplant clade species might originate from a single endogenization event that would have happened between the late Pliocene and the early Pleistocene. This hypothesis now needs to be statistically tested by using a much more broadly sampled phylogenetic framework that will allow to confirm whether BBVA insertion happened several times

in the giant genus *Solanum* or if it is a true synapomorphy of the Eggplant clade.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

SS conceived the study, performed the analyses, and analyzed the data with help from FM, AG, and XA. VS, SS, and P-YT drafted the manuscript. All authors contributed to the final version.

FUNDING

This study was funded by the Agence Nationale de la Recherche (ANR) grant ANR-17-CE20-0001 EVENTS.

ACKNOWLEDGMENTS

Authors thank Marie-Christine Daunay (INRAE Montfavet, France) for generously providing eggplant seeds.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.683681/full#supplementary-material

REFERENCES

Aubriot, X., Knapp, S., Syfert, M. M., Poczai, P., and Buerki, S. (2018). Shedding new light on the origin and spread of the brinjal eggplant (Solanum melongena L.) and its wild relatives. *Am. J. Bot.* 105, 1175–1187. doi: 10.1002/ajb2.1133

Aubriot, X., Singh, P., and Knapp, S. (2016). Tropical Asian species show that the old world clade of "spiny solanums" (Solanum subgenus Leptostemonum pro parte: Solanaceae) is not monophyletic. *Bot. J. Linn. Soc.* 181, 199–223. doi: 10.1111/boj.12412

Ballvora, A., Ercolano, M. R., Weiß, J., Meksem, K., Bormann, C. A., Oberhagemann, P., et al. (2002). The R1 gene for potato resistance to late blight (*Phytophthora infestans*) belongs to the leucine zipper/NBS/LRR class of plant resistance genes. *Plant J.* 30, 361–371. doi: 10.1046/j.1365-313X.2001.01292.x

Barchi, L., Pietrella, M., Venturini, L., Minio, A., Toppino, L., Acquadro, A., et al. (2019). A chromosome-anchored eggplant genome sequence reveals key events in *Solanaceae* evolution. *Sci. Rep.* 9:11769. doi: 10.1038/s41598-019-47985-w

Bennetzen, J. L. (2000). Transposable element contributions to plant gene and genome evolution. *Plant Mol. Biol.* 42, 251–269. doi: 10.1023/A:1006344508454

Bertsch, C., Beuve, M., Dolja, V. V., Wirth, M., Pelsy, F., Herrbach, E., et al. (2009). Retention of the virus-derived sequences in the nuclear genome of grapevine as a potential pathway to virus resistance. *Biol. Direct* 4:21. doi: 10.1186/1745-6150-4-21

Beyer, U., Moll-Rocek, J., Moll, U. M., and Dobbelstein, M. (2011). Endogenous retrovirus drives hitherto unknown proapoptotic p63 isoforms in the male germ line of humans and great apes. *Proc. Natl. Acad. Sci. U.S.A.* 108, 3624–3629. doi: 10.1073/pnas.1016201108

Carrasco, J. L., Sánchez-Navarro, J. A., and Elena, S. F. (2019). Exploring the role of cellular homologous of the 30K-superfamily of plant virus movement proteins. *Virus Res.* 262, 54–61. doi: 10.1016/j.virusres.2018.02.015

Chabannes, M., Baurens, F.-C., Duroy, P.-O., Bocs, S., Vernerey, M. S., Goud, M. R., et al. (2013). Three infectious viral species lying in wait in the banana genome. *J. Virol.* 87, 8624–8637. doi: 10.1128/jvi.00899-13

Cosby, R. L., Judd, J., Zhang, R., Zhong, A., Garry, N., Pritham, E. J., et al. (2021). Recurrent evolution of vertebrate transcription factors by transposase capture. *Science* 371:eabc6405. doi: 10.1126/science.abc6405

Diop, S. I., Geering, A. D. W., Alfama-Depauw, F., Loaec, M., Teycheney, P. Y., and Maumus, F. (2018). Tracheophyte genomes keep track of the deep evolution of the Caulimoviridae. *Sci. Rep.* 8:572. doi: 10.1038/s41598-017-16399-x

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635

Dupressoir, A., Marceau, G., Vernochet, C., Bénit, L., Kanellopoulos, C., Sapin, V., et al. (2005). Syncytin-A and syncytin-B, two fusogenic placenta-specific murine envelope genes of retroviral origin conserved in Muridae. *Proc. Natl. Acad. Sci. U.S.A.* 102, 725–730. doi: 10.1073/pnas.0406509102

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461

Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* 9:18. doi: 10.1186/1471-2105-9-18

Faino, L., Carli, P., Testa, A., Cristinzio, G., Frusciante, L., and Ercolano, M. R. (2010). Potato R1 resistance gene confers resistance against Phytophthora infestans in transgenic tomato plants. *Eur. J. Plant Pathol.* 128, 233–241. doi: 10.1007/s10658-010-9649-2

Faulkner, G. J., Kimura, Y., Daub, C. O., Wani, S., Plessy, C., Irvine, K. M., et al. (2009). The regulated retrotransposon transcriptome of mammalian cells. *Nat. Genet.* 41, 563–571. doi: 10.1038/ng.368

Feschotte, C., and Gilbert, C. (2012). Endogenous viruses: insights into viral evolution and impact on host biology. *Nat. Rev. Genet.* 13, 283–296. doi: 10.1038/nrg3199

Filloux, D., Murrell, S., Koohapitagtam, M., Golden, M., Julian, C., Galzi, S., et al. (2015). The genomes of many yam species contain transcriptionally active endogenous geminiviral sequences that may be functionally expressed. *Virus Evol.* 1:vev002. doi: 10.1093/ve/vev002

Fry, W. E., and Goodwin, S. B. (1997). Resurgence of the Irish potato famine fungus: after 150 years, the late blight fungus is again menacing farmers. *Bioscience* 47, 363–371. doi: 10.2307/1313151

Fütterer, J., Potrykus, I., Valles Brau, M. P., Dasgupta, I., Hull, R., and Hohn, T. (1994). Splicing in a plant pararetrovirus. *Virology* 198, 663–670.

Geering, A. D. W., Maumus, F., Copetti, D., Choisne, N., Zwickl, D. J., Zytnicki, M., et al. (2014). Endogenous florendoviruses are major components of plant genomes and hallmarks of virus evolution. *Nat. Commun.* 5:5269. doi: 10.1038/ncomms6269

Geuking, M. B., Weber, J., Dewannieux, M., Gorelik, E., Heidmann, T., Hengartner, H., et al. (2009). Recombination of retrotransposon and exogenous RNA virus results in nonretroviral cDNA integration. *Science* 323, 393–396. doi: 10.1126/science.1167375

Goff, L. A., Trapnell, C., and Kelley, D. (2012). *CummeRbund: Visualization And Exploration Of Cufflinks High-Throughput Sequencing Data*. Available online at: http://compbio.mit.edu/cummeRbund/manual/cummeRbund-manual.html

Gong, Z., and Han, G.-Z. (2018). Euphyllophyte paleoviruses illuminate hidden diversity and macroevolutionary mode of caulimoviridae. *J. Virol* 92, JVI.2043–JVI.2017. doi: 10.1128/jvi.02043-17

Goss, E. M., Tabima, J. F., Cooke, D. E. L., Restrepo, S., Frye, W. E., Forbes, G. A., et al. (2014). The Irish potato famine pathogen Phytophthora infestans originated in central Mexico rather than the Andes. *Proc. Natl. Acad. Sci. U.S.A.* 111, 8791–8796. doi: 10.1073/pnas.1401884111

Grünwald, N. J., and Flier, W. G. (2005). The biology of Phytophthora infestans at its center of origin. *Annu. Rev. Phytopathol.* 43, 171–190. doi: 10.1146/annurev.phyto.43.040204.135906

Harrach, B., Both, G. W., Brown, M., Davison, A. J., and Echavarria, M. (2011). *Virus Taxonomy. In: Family Adenoviridae: Classification and Nomenclature of Viruses. Ninth Report of the International Committee on Taxonomy of Viruses*. Elsevier: Academic Press, San Diego.

Holmes, E. C. (2011). The evolution of endogenous viral elements. *Cell Host Microbe* 10, 368–377. doi: 10.1016/j.chom.2011.09.002

Horie, M., Honda, T., Suzuki, Y., Kobayashi, Y., Daito, T., Oshida, T., et al. (2010). Endogenous non-retroviral RNA virus elements in mammalian genomes. *Nature* 463, 84–87. doi: 10.1038/nature08695

Huelsenbeck, J. P., and Ronquist, F. (2001). MRBAYES: bayesian inference of phylogenetic trees. *Bioinformatics* 17, 754–755. doi: 10.1093/bioinformatics/17.8.754

Hull, R. (2009). *Comparative Plant Virology. 2nd ed*. Massachusetts: Academic Press Inc.

Knapp, S., Vorontsova, M. S., and Prohens, J. (2013). Wild relatives of the eggplant (Solanum melongena L.: Solanaceae): new understanding of species names in a complex group. *PLoS One* 8:e57039. doi: 10.1371/journal.pone.0057039

Kuang, H., Wei, F., Marano, M. R., Wirtz, U., Wang, X., Liu, J., et al. (2005). The R1 resistance gene cluster contains three groups of independently evolving, type I R1 homologues and shows substantial structural variation among haplotypes of *Solanum demissum*. *Plant J.* 44, 37–51. doi: 10.1111/j.1365-313X.2005.02506.x

Kuriyama, K., Tabara, M., Moriyama, H., Kanazawa, A., Koiwa, H., Takahashi, H., et al. (2020). Disturbance of floral colour pattern by activation of an endogenous pararetrovirus, petunia vein clearing virus, in aged petunia plants. *Plant J.* 103, 497–511. doi: 10.1111/tpj.14728

Leebens-Mack, J. H., Barker, M. S., Carpenter, E. J., Deyholos, M. K., Gitzendanner, M. A., Graham, S. W., et al. (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574, 679–685. doi: 10.1038/s41586-019-1693-2

Lescot, M., Déhais, P., Thijs, G., Marchal, K., Moreau, Y., Van De Peer, Y., et al. (2002). PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for *in silico* analysis of promoter sequences. *Nucleic Acids Res.* 30, 325–327. doi: 10.1093/nar/30.1.325

Lockhart, B. E., Menke, J., Dahal, G., and 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, 1579–1585. doi: 10.1099/0022-1317-81-6-1579

Lu, S., Wang, J., Chitsaz, F., Derbyshire, M. K., Geer, R. C., Gonzales, N. R., et al. (2020). CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res.* 48, D265–D268. doi: 10.1093/nar/gkz991

Martin, A., Troadec, C., Boualem, A., Rajab, M., Fernandez, R., Morin, H., et al. (2009). A transposon-induced epigenetic change leads to sex determination in melon. *Nature* 461, 1135–1138. doi: 10.1038/nature08498

Martin, D. P., Murrell, B., Golden, M., Khoosal, A., and Muhire, B. (2015). RDP4: detection and analysis of recombination patterns in virus genomes. *Virus Evol.* 1:vev003. doi: 10.1093/ve/vev003

Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S. L., et al. (2013). FUBAR: a fast, unconstrained bayesian AppRoximation for inferring selection. *Mol. Biol. Evol.* 30, 1196–1205. doi: 10.1093/molbev/mst030

Mushegian, A. R., and Elena, S. F. (2015). Evolution of plant virus movement proteins from the 30K superfamily and of their homologs integrated in plant genomes. *Virology* 476, 304–315. doi: 10.1016/j.virol.2014.12.012

Mette, M. F., Kanno, T., Aufsatz, W., Jakowitsch, J., Winden, J., Van Der, Matzke, M. A., et al. (2002). Endogenous viral sequences and their potential contribution to heritable virus resistance in plants. *EMBO J.* 21, 461–469. doi: 10.1093/emboj/21.3.461

Meyer, T. J., Rosenkrantz, J. L., Carbone, L., and Chavez, S. L. (2017). Endogenous retroviruses: with us and against us. *Front. Chem.* 5:23. doi: 10.3389/fchem.2017.00023

Min, X. J. (2013). Asfinder: a tool for genome-wide identification of alternatively splicing transcripts from EST-derived sequences. *Int. J. Bioinform. Res. Appl.* 9, 221–226. doi: 10.1504/IJBRA.2013.053603

Mitreiter, K., Schmidt, J., Luz, A., Atkinson, M. J., Höfler, H., Erfle, V., et al. (1994). Disruption of the murine p53 gene by insertion of an endogenous retrovirus-like element (ETn) in a cell line from radiation-induced osteosarcoma. *Virology* 200, 837–841. doi: 10.1006/viro.1994.1253

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

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Novák, P., Neumann, P., Pech, J., Steinhaisl, J., and MacAs, J. (2013). RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics* 29, 792–793. doi: 10.1093/bioinformatics/btt054

Nylander, J. A. A., Ronquist, F., Huelsenbeck, J. P., and Nieves-Aldrey, J. L. (2004). Bayesian phylogenetic analysis of combined data. *Syst. Biol.* 53, 47–67. doi: 10.1080/10635150490264699

Page, A., Gibson, J., Meyer, R. S., and Chapman, M. A. (2019). Eggplant domestication: pervasive gene flow, feralization, and transcriptomic divergence. *Mol. Biol. Evol.* 36, 1359–1372. doi: 10.1093/molbev/msz062

Rambaut, A., Suchard, M. A., Xie, D., and Drummond, A. J. (2014). *Tracer v1. 6. Computer program and Documentation Distributed By The Author*. Available online at: http://beast.bio.ed.ac.uk/Tracer. http://beast.bio.ed.ac.uk/tracer accussed on Mar 26, 2021

Reese, M. G. (2001). Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome. *Comput. Chem.* 26, 51–56. doi: 10.1016/S0097-8485(01)00099-7

Richert-Pöggeler, K. R., Noreen, F., Schwarzacher, T., Harper, G., and Hohn, T. (2003). Induction of infectious petunia vein clearing (pararetro) virus from endogenous provirus in petunia. *EMBO J.* 22, 4836–4845. doi: 10.1093/emboj/cdg443

Sanseverino, W., Roma, G., De Simone, M., Faino, L., Melito, S., Stupka, E., et al. (2009). PRGdb: a bioinformatics platform for plant resistance gene analysis. *Nucleic Acids Res.* 38, D814–D821. doi: 10.1093/nar/gkp978

Sha, M., Lee, X., Li, X., Veldman, G. M., Finnerty, H., Racie, L., et al. (2000). Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* 403, 785–789. doi: 10.1038/35001608

Shahmuradov, I. A., Umarov, R. K., and Solovyev, V. V. (2017). TSSPlant: a new tool for prediction of plant Pol II promoters. *Nucleic Acids Res.* 45:e65. doi: 10.1093/nar/gkw1353

Sharma, V., Lefeuvre, P., Roumagnac, P., Filloux, D., Teycheney, P.-Y., Martin, D. P., et al. (2020). Large-scale survey reveals pervasiveness and potential function of endogenous geminiviral sequences in plants. *Virus Evol.* 6:veaa071. doi: 10.1093/ve/veaa071

Song, B., Song, Y., Fu, Y., Kizito, E. B., Kamenya, S. N., Kabod, P. N., et al. (2019). Draft genome sequence of Solanum aethiopicum provides insights into disease resistance, drought tolerance, and the evolution of the genome. *Gigascience* 8, 1–16. doi: 10.1093/gigascience/giz115

Sullivan, M. J., Petty, N. K., and Beatson, S. A. (2011). Easyfig: a genome comparison visualizer. *Bioinformatics* 27, 1009–1010. doi: 10.1093/bioinformatics/btr039

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 34, W609–W612. doi: 10.1093/nar/gkl315

Taylor, D. J., and Bruenn, J. (2009). The evolution of novel fungal genes from non-retroviral RNA viruses. *BMC Biol.* 7:88. doi: 10.1186/1741-7007-7-88

Team R Development Core. (2018). *A Language and Environment for Statistical Computing. R Found. Stat. Comput. 2*. Available online at: https://www.R-project.org. Available online at: http://www.r-project.org. acussed date May 18 2021

Teycheney, P., and Geering, A. (2011). "Endogenous viral sequences in plant genomes," in *Recent Advances in Plant Virology*, eds M. Tepfer, J. J. Lopez-Moya, C. Caranta, and M. A. Aranda (Caister: Academic Press), 343–362.

Teycheney, P. Y., Geering, A. D. W., Dasgupta, I., Hull, R., Kreuze, J. F., Lockhart, B., et al. (2020). ICTV Virus taxonomy profile: caulimoviridae. *J. Gen. Virol.* 101, 1025–1026. doi: 10.1099/jgv.0.001497

Teycheney, P. Y., and Tepper, M. (2007). Possible roles of endogenous plant viral sequences and transgenes containing viral sequences in both virus resistance and virus emergence. *Environ. Biosafety Res.* 6, 219–221. doi: 10.1051/ebr:2007045

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562–578. doi: 10.1038/nprot.2012.016

Weaver, S., Shank, S. D., Spielman, S. J., Li, M., Muse, S. V., and Kosakovsky Pond, S. L. (2018). Datamonkey 2.0: a modern web application for characterizing selective and other evolutionary processes. *Mol. Biol. Evol.* 35, 773–777. doi: 10.1093/molbev/msx335

Zeng, F., and Cheng, B. (2014). Transposable element insertion and epigenetic modification cause the multiallelic variation in the expression of FAE1 in Sinapis alba. *Plant Cell* 26, 2648–2659. doi: 10.1105/tpc.114.126631

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.