Ciência Agrícola, Rio Largo, v. 17, n. 2, p. 59-72, 2019

DIVERSIFICATION INTO THE GENUS *BADNAVIRUS*: PHYLOGENY AND POPULATION GENETIC VARIABILITY

Caio Henrique Loureiro de Holanda Ferreira[1], Lucas Jobim Jordão[1], Roberto Ramos-Sobrinho[2], Mayra Machado de Medeiros Ferro[1*], Sarah Jacqueline Cavalcanti da Silva[1], Iraildes Pereira Assunção[1], Gaus Silvestre de Andrade Lima[1]

[1]Setor de Fitossanidade/Centro de Ciências Agrárias, Universidade Federal de Alagoas, Rio Largo, AL, Brazil, 57100-000
[2]School of Plant Sciences, University of Arizona, Tucson, AZ, 85721, USA
***Autor para correspondência:** Mayra Machado de Medeiros Ferro, mayra.ferro@hotmail.com

ABSTRACT: Badnaviruses (family *Caulimoviridae) have semicircular dsDNA genomes encapsidated into bacilliform particles. The genus Badnavirus is the most important due to its high number of species reported infecting* cultivated plants worldwide. This study aimed to evaluate the phylogenetic positioning and population genetic variability into *Badnavirus. Data sets comprising the badnavirus complete genome and partial sequences of the RT and RNaseH genes were obtained from the GenBank database.* Multiple nucleotide sequence alignments from complete genome, ORFIII, complete genomic domain RT/RNaseH (1020pb) and partial (579pb) were performed. A total of 127 genomes were obtained, representing 53 species of badnavirus. Nucleotide sequence comparisons for the RT/RNaseH domain showed only a few isolates reported as distinct species shared ≥80% identity, the current threshold used for species demarcation into this genus. Phylogenetic trees for the complete genome and for ORFIII showed four well supported clusters (badnavirus groups 1-4), with clusters 1 and 3 being sister groups comprising predominantly sugarcane- and banana-infecting species. Non-tree-like evolution analysis evidenced putative recombination events among badnaviruses, and at least 23 independent events were detected. High levels of nucleotide diversity were observed for the partial RT/RNaseH region in isolates of 11 badnavirus species. These results showed that mutation and recombination are important mechanisms that acting on badnavirus diversification.

KEY WORDS: evolution, recombination, species demarcation.

DIVERSIFICAÇÃO DENTRO DO GÊNERO *Badnavirus*: FILOGENIA E VARIABILIDADE GENÉTICA DA POPULAÇÃO

RESUMO: Badnaviroses (família Caulimoviridae) possuem genomas semicirculares de dsDNA encapsidados em partículas baciliformes. O gênero *Badnavirus* é o mais importante devido ao seu alto número de espécies relatadas infectando plantas cultivadas em todo o mundo. Este estudo teve como objetivo avaliar o posicionamento filogenético e a variabilidade genética populacional em *Badnavirus*. Conjuntos de dados compreendendo genoma completo de badnavirus e sequências parciais de RT/RNaseH foram recuperados do GenBank. Foram obtidos alinhamentos múltiplos de sequências nucleotídicas para o genoma completo, ORFIII, domínio RT/RNaseH completo (1020pb) e parcial (579pb). Um total de 127 genomas foram obtidos, representando 53 espécies de badnavírus. Comparações de sequências nucleotídicas para RT/RNaseH mostraram alguns isolados relatados como espécies distintas compartilhando ≥80% de identidade, limiar atual usado para demarcação de espécies para este gênero. As árvores filogenéticas para o genoma completo e ORFIII apresentaram quatro grupos bem suportados (grupos de badnavírus 1-4), sendo os grupos 1 e 3 os grupos-irmãos, compostos predominantemente por espécies que infectam cana-de-açúcar e banana. Evolução em rede evidenciou eventos putativos de recombinação entre os badnavírus, e pelo menos 23 eventos independentes foram detectados. Altos níveis de diversidade nucleotídica foram observados para a região de RT/RNaseH parcial em isolados de 11 espécies de badnavírus. Estes resultados mostraram que mutação e recombinação são mecanismos importantes que atuam na diversificação de badnavírus.

PALAVRAS CHAVE: evolução, recombinação, demarcação de espécies.

INTRODUCTION

Viruses belonging to the family *Caulimoviridae* have semicircular, double-stranded (ds)DNA genomes, 7.2-9.2 kbp in length, encapsidated into isometric or bacilliform particles, and which replicate through an RNA intermediate (plant pararetroviruses; Geering and Hull, 2012). This family is divided into eight genera (*Badnavirus, Caulimovirus*, *Cavemovirus*, *Petuvirus, Rosadnavirus, Solendovirus, Soymovirus* and *Tungrovirus*) according to host range, insect vector, genomic organization and phylogeny (Geering and Hull, 2012; Bath et al., 2016). Badnaviruses are transmitted mostly by mealybugs (a few species by aphids) in a semi-persistent manner (Geering and Hull, 2012; Bath et al., 2016) and are among the most important plant viruses with a DNA genome.

The semicircular dsDNA of badnaviruses has site-specific discontinuities and at least three ORFs, named I, II and III (Bouhida et al., 1993; Hagen et al., 1993; Harper and Hull, 1998; Geering and Hull, 2012). Proteins codified by ORFs I and II have been reported to be virion-associated (Cheng et al., 1996) and nucleic acid-binding (ORF II; Jacquot et al., 1996). ORFIII encodes a polyprotein of 208-216 kDa that is proteolytically cleaved generating the movement and coat proteins, the aspartate protease responsible for the polyprotein cleavage, the reverse transcriptase (RT) and ribonuclease H (RNaseH) (both genomic domains involved in the viral replication) (Medberry et al., 1990; Harper and Hull, 1998). The criterion of ≥80% nucleotide sequence identity for the RT/RNaseH domains was established for species demarcation in the genus *Badnavirus* (Geering and Hull, 2012), and specific primer pairs are largely used to amplify this viral genomic region (Yang et al., 2003). However, different studies have shown this criterion to be insufficient to separate some badnavirus species, mostly those infecting banana and sugarcane (Muller et al., 2011; Karuppaiah et al., 2013; Silva et al., 2015). Furthermore, the existence of endogenous badnavirus sequences represents a great challenge for taxonomy and diagnosis of members into this genus.

Different badnavirus species have been reported infecting economically important crops such as sugarcane (*Saccharum* spp.), banana (*Musa* spp.), cacao (*Theobroma cacao* L.), black pepper (*Piper* spp.), yam (*Dioscorea* spp.) and citrus (*Citrus* spp.) (Eni et al., 2008; James et al., 2011; Johnson et al., 2012; Kouakou et al., 2012; Deeshma and Bhat, 2015; Silva et al., 2015), with 40 badnavirus species being currently recognized by the International Committee on Taxonomy of Viruses (https://talk.ictvonline.org/taxonomy/). Badnaviruses have also been reported as integrated sequences into their host genomes, commonly referred to as endogenous pararetroviruses (EPRVs; Hohn et al., 2008; Staginnus et al., 2009), and it is known some EPRVs are able to 'scape' from the host genome and establish a systemic infection (episomal forms; Dallot et al., 2001; Lheureux et al., 2003; Côte et al., 2010). Besides this high species diversity, badnavirus populations also show high levels of molecular variability (Geering et al., 2000; Geijskes et al., 2002; Harper et al., 2005; Govind et al., 2014; Guimarães et al., 2015; Silva et al., 2015). The lack of a proofreading activity of the badnavirus reverse transcriptase (Svarovskaia et al., 2003) and the occurrence of recombination events (Govind et al., 2014) may directly affect the high genetic variability observed in this viral group.

A *in silico* large-scale study was carried out to obtain more information about the genetic relationship and variability in *Badnavirus*. The threshold for nucleotide sequence comparisons of the RT/RNaseH genomic region currently used for species demarcation identifies most reported badnaviruses. However, this criterion alone is unable to differentiate all sugarcane- and banana-infecting badnaviruses. A new badnavirus phylogenetic clade was proposed, here named badnavirus group 4. Additionally, it was observed this is a highly diverse viral group, with recombination and mutation being important factors contributing to high levels of nucleotide diversity observed in some badnavirus populations.

MATERIAL AND METHODS

Badnavirus data set

Full-length genome sequences of badnaviruses were retrieved from the non-redundant GenBank database (www.ncbi.nlm.nih.gov/genbank; accessed on Nov 2017) (Table 1). Data sets of nucleotide sequences of the ORFIII, and full (1020pb) and partial (579pb) RT/RNaseH domains were obtained from the complete genomes. The partial RT/RNaseH data set comprises the genomic region amplified by the primer pairs largely used for detection and identification of badnaviruses (Yang et al., 2003). A previous analysis using sequences from

ORF I and II showed these regions are inconclusive, presenting low support and insufficient phylogenetic signal, for this reason were not included in this study (data not shown).

Table 1. Full-length badnavirus sequences retrieved from the non-redundant GenBank database on Nov 2017.

Species/Acronym	GenBank access#
Banana streak CA virus/BSCAV	HQ593111, KJ013511
Banana streak GF virus/BSGFV	AY493509, KJ013507
Banana streak IM virus/BSIMV	HQ593112, KJ013508, HQ659760, KT895258
Banana streak MY virus/BSMYV	KR014107, KJ013509, KF724855, KF724856, AY805074, KF724854
Banana streak OL virus/BSOLV	JQ409540, JQ409539, KJ013506, DQ451009, DQ859899
Banana streak UA virus/BSUAV	HQ593107
Banana streak UI virus/BSUIV	HQ593108
Banana streak UL virus/BSULV	HQ593109
Banana streak UM virus/BSUMV	HQ593110
Banana streak virus Acuminata Yunnan/BSVNV	DQ092436, AY750155, KJ013510
Blackberry virus F/BVF	KJ413252
Bougainvillea spectabilis chlorotic vein-banding virus/BCVBV	EU034539
Cacao mild mosaic virus/CaMMV	KX276640
Cacao swollen shoot CD virus /CSSCDV	JN606110
Cacao swollen shoot Togo A virus/CSSTAV	AJ781003
Cacao red vein virus/CRVV	KX592584, KX592573, KX592572, KX592571,
Cacao swollen shoot virus/CSSV	KX592582, KX592583, KX592581, KX592580, KX592579, KX592578, KX592577, KX592575, KX592574, AJ609019, AJ534983, L14546
Cacao yellow vein-banding virus /CYVBV	KX276641
Canna yellow mottle virus/CaYMV	KU168312, MF074075
Canna yellow mottle-associated virus/CaYMAV	KX066020
Citrus yellow mosaic virus/CYMV	EU489745, EU489744, EU708316, JN006806, JN006805, EU708317, AF347695
Commelina yellow mottle virus/ CoYMV	X52938
Cycad leaf necrosis virus/CYLNV	EU853709
Dioscorea bacilliform AL virus/DBALV	KX008573, KX008572, KX008571
Dioscorea bacilliform RT virus /DBRTV	KX430257
Dioscorea bacilliform RT virus 1 /DBRTV1	KX008576, KX008575, KX008574
Dioscorea bacilliform RT virus 2/DBRTV2	KX008579, KX008578, KX008577, KY827393
Dioscorea bacilliform SN virus/DBSNV	DQ822073
Dioscorea bacilliform AL virus 2 /DBALV2	KY827395
Dioscorea baciliform ES virus/DBESV	KY827394
Dracaena mottle virus/DMV	DQ473478, EF494181
Fig badnavirus 1/FBV-1	KT809307, KT809306, KT809305, KT809304, KT809303, JF411989
Gooseberry vein banding virus/GVBV	HQ852251, HQ852250, HQ852249, HQ852248, JQ316114
Grapevine roditis leaf discoloration-associated virus/GRLDaV	HG940503, KT965859
Grapevine vein-clearing virus/GVCV	KX610317, KX610316, KT907478, JF301669, KJ725346
Hibiscus bacilliform virus/HBV	KF875586
Kalanchoe top-spotting virus/KTSV	AY180137
Mulberry badnavirus 1/MBV-1	LN651258
Pagoda yellow mosaic associated virus /PYMAV	KJ013302
Pineapple bacilliform comosus virus/PBCV	GU121676, GQ398110
Piper yellow mottle virus/PYMoV	KJ873043, KJ873041, KJ873042, KC808712
Rubus yellow net virus/RYNV	KF241951, KM078034
Sugarcane bacilliform GA virus/SCBGAV	FJ824813, FJ824814, KJ624754
SCBGAV/BSV	KT895259
Sugarcane bacilliform GD virus/SCBGDV	FJ439817
Sugarcane bacilliform IM virus/SCBIMV	AJ277091, JN377537, JN377536, JN377533, KM214358
Sugarcane bacilliform MO virus/SCBMOV	M89923, JN377534
Sugarcane bacilliform BB virus/SCBBBV	KM214357, JN377535
Sweet potato pakakuy virus/SPPV	FJ560943
Sweet potato badnavirus B/SPBV	FJ560944
Taro bacilliform CH virus/TaBCHV	KP710178, KP710177
Taro bacilliform virus/TaBV	AF357836
Wisteria badnavirus 1/WBV-1	KX168422
Yacon necrotic mottle virus /YNMoV	KM229702

Sequence analysis

Multiple amino acid sequence alignments were prepared for the ORFIII using the MUSCLE algorithm (Edgar, 2004), manually edited in the MEGA7 package (Kumar et al., 2016) and returned to nucleotide sequences for posterior analyses. The RT/RNaseH sequences (full and partial) were obtained from the ORFIII data set. Additionally, multiple nucleotide alignments were obtained for the complete genome. To confirm taxonomy attributed to the badnavirus isolates retrieved from GenBank, pairwise nucleotide sequence comparisons were performed to all data sets using Sequence Demarcation Tool (SDT) v.1.2 (Muhire et al., 2013).

Phylogenetic inference

In order, to demonstrate if the phylogenetic relationship observed for the RT/RNaseH reflects the clustering inferred for the complete genomes of badnaviruses, Bayesian phylogenetic trees were obtained for the complete genome, ORFIII and RT/RNaseH (full and partial) data sets. Analyses were run using MrBayes v. 3.2 (Ronquist et al., 2012) through the CIPRES web portal (Miller et al., 2010), assuming GTR+I+G as the evolutionary model. Two replicates with four chains each for 20 million generations and sampling every 2,000 generations were used. The first 2,500 trees were discarded as a burn-in phase in each run. Posterior probabilities (Rannala and Yang, 1996) were determined from a majority-rule consensus tree generated with the 15,000 remaining trees. The trees were edited in FigTree v.1.4 (ztree.bio.ed.ac.uk/software/figtree) and Inkscape (https://inkscape.org/pt/).

Recombination analysis

Evidence of non-tree-like evolution was assessed for the complete genome, ORFIII, and RT/RNaseH (full and partial) data sets using the Neighbor-Net method implemented in SplitsTree v.4.10 (Huson and Bryant, 2006). Putative parental sequences and recombination breakpoints for the complete genome data set were determined using the methods RDP, Geneconv, Boot-scan, Maximum Chi Square, Chimaera, SisterScan and 3Seq implemented in the RDP v.4.0 package (Martin et al., 2015). Alignments were analyzed with default settings for the different methods and statistical significance was inferred by a *P*-value lower than a Bonferroni-corrected cut-off of 0.05. Only events detected by at least five different methods were considered to be reliable.

Population genetic variability

Partial nucleotide sequences of the RT/RNaseH region of badnaviruses infecting different hosts were retrieved from the non-redundant GenBank database (www.ncbi.nlm.nih.gov/genbank; accessed on Dez 2017). The mean pairwise number of nucleotide differences per site (nucleotide diversity, π) was estimated for each population using DnaSP v. 6.10 (Rozas et al., 2017).

RESULTS

Badnavirus isolates

A total of 127 full-length genomes were obtained from GenBank (Table 1), comprising 53 different badnavirus species. *Cacao swollen shoot virus* (CSSV) was the badnavirus represented by the higher number of sequences/isolates (12), while 25 species were represented by one only sequence (Table 1).

Species demarcation

Pairwise sequence comparisons for the partial (579pb) and full (1020pb) RT/RNaseH, largely used for badnavirus species identification, showed percent nucleotide identities ranging from 57.2 to 82.4% and 57.1 to 83.8%, respectively, among species. Therefore, some comparisons exceeded the currently used 80% nucleotide identity criterion for species demarcation into the genus *Badnavirus* (Geering and Hull, 2012).

According to the partial RT/RNaseH region, *Sugarcane bacilliform GA virus* (SCBGAV) isolates shared 80.5-82.0% and 81.3-82.4% identity with *Banana streak OL virus* (BSOLV) and *Banana streak CA virus* (BSCAV) isolates, respectively, with BSOLV and BSCAV showing up to 77.7% identity. *Sugarcane bacilliform BB virus* (SCBBBV) showed 77.4-80.4% and 78.3-82.0% nucleotide identity with *Sugarcane bacilliform IM virus* (SCBIMV) and *Sugarcane bacilliform MO virus* (SCBMOV), respectively, while SCBIMV and SCBMOV isolates shared 77,9-79,8%

identity. To the full RT/RNaseH, the percentages of nucleotide identity were slightly increased: SCBGAV shared 81.0-82.4% and 81.9-82.5% identity with BSOLV and BSCAV, respectively, with BSOLV and BSCAV showing up to 79.0% identity. SCBBBV showed 78.7-81.2% and 78.8-81.9% identity with SCBIMV and SCBMOV, respectively, while SCBIMV and SCBMOV isolates shared 78.4-80.9% identity.

The *Sweet potato badnavirus B* (SPBV) and *Sweet potato pakakuy virus* (SPPV) isolates shared 81.9% and 83.8% of nucleotide identity for the partial and full RT/RNaseH data sets, respectively. For all badnavirus represented by more than one sequence/ isolate, percent nucleotide identities were higher than 80.0% within species.

When analyzed the ORFIII data set, which comprises the RT/RNaseH domains, SCBGAV showed highest nucleotide identity of 80.3% and 79.8% with BSOLV and BSCAV, respectively. However, SCBBBV, SCBIMV and SCBMOV shared up to 75,7% identity. For the complete genome, all pairwise comparisons between distinct badnaviruses were lower than 80% identity, with SCBGAV isolates showing the highest level of nucleotide identity (79.5%) with BSOLV and BSCAV isolates.

Phylogenetic relationship

In the complete genome Bayesian phylogenetic tree, the three badnavirus clusters (badnavirus groups 1, 2 and 3) described by Muller et al., (2011) were observed (Figure 1). Additionally, a fourth clade can be observed, here named badnavirus group 4 (Figure 1). The clusters 1 and 3 formed sister groups, being predominantly comprised by badnaviruses infecting sugarcane and banana (Figure 1). Similar results were observed for the ORFIII data set (Figure 2), which represents ~80% of the complete genome. These results reinforce the idea that badnaviruses infecting sugarcane and banana are closely related, as indicated by the pairwise comparisons and recombination analyses (see below).

When analyzed the phylogenetic trees for the RT/RnaseH data sets (full and partial), badnavirus groups are still evidenced, but with many topological incongruences and very low resolution (SFigure 1 and SFigure 2). Some species (mainly the cacao-infecting isolates) in the badnavirus group 2 clustered with isolates in groups 3 and 4, while the other group 2 isolates clustered with badnaviruses in group 1 (SFigure 1). However, statistical support was considerably smaller for trees based on RT/ RnaseH sequences than ORFIII and complete genome, indicating the RT/RNaseH sequences have insufficient phylogenetic signal.

Recombination events

Non-tree-like evolution for the complete genome, ORF III and RT/RNaseH revealed evidence of putative recombination events affecting the evolution of badnaviruses, with both intra- and interspecies recombination being observed (Figure 3, Figure 4, SFigure 3 and SFigure 4). These events were more pronounced among badnaviruses infecting sugarcane and banana (phylogenetic groups 1 and 3), and cacao (phylogenetic group 2). Besides recombination, a strong mutation effect was evidenced in the diversification of members into this genus, indicated by the long branches associated with many isolates (Figure 3, Figure 4, SFigure 3 and SFigure 4).

Figure 1. Midpoint-rooted Bayesian phylogenetic trees based on complete genome nucleotide sequences of badnaviruses, indicating badnavirus groups 1 (in grey), 2 (green), 3 (blue) and 4 (orange). Commelina yellow mottle virus (CoYMV; in black) was placed apart from the four main clusters.

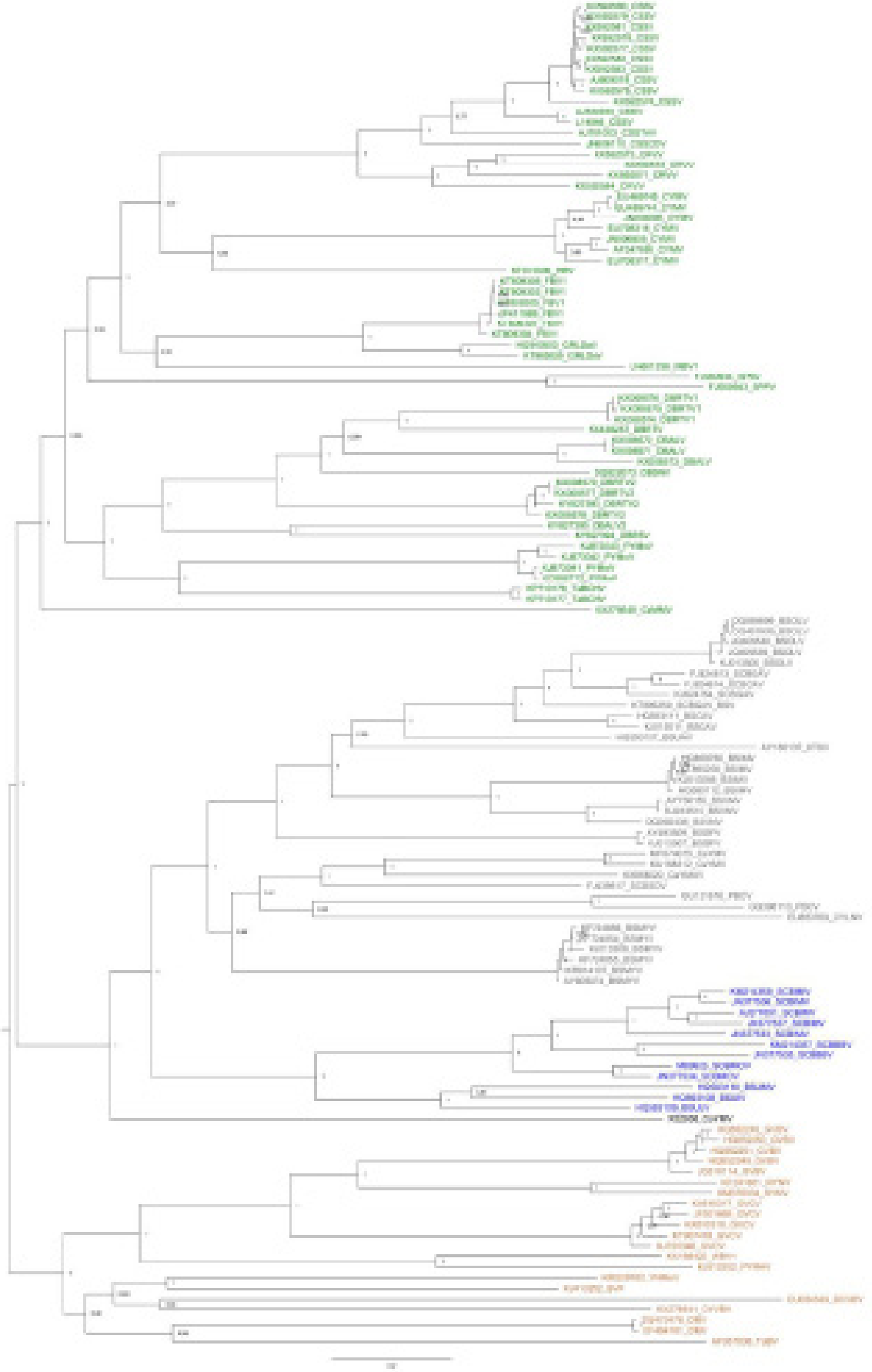


Figure 2. Midpoint-rooted Bayesian phylogenetic trees based on the ORF III nucleotide sequences of badnaviruses, indicating badnavirus groups 1 (in grey), 2 (green), 3 (blue) and 4 (orange). Commelina yellow mottle virus (CoYMV; in black) was placed apart from the four main clusters.

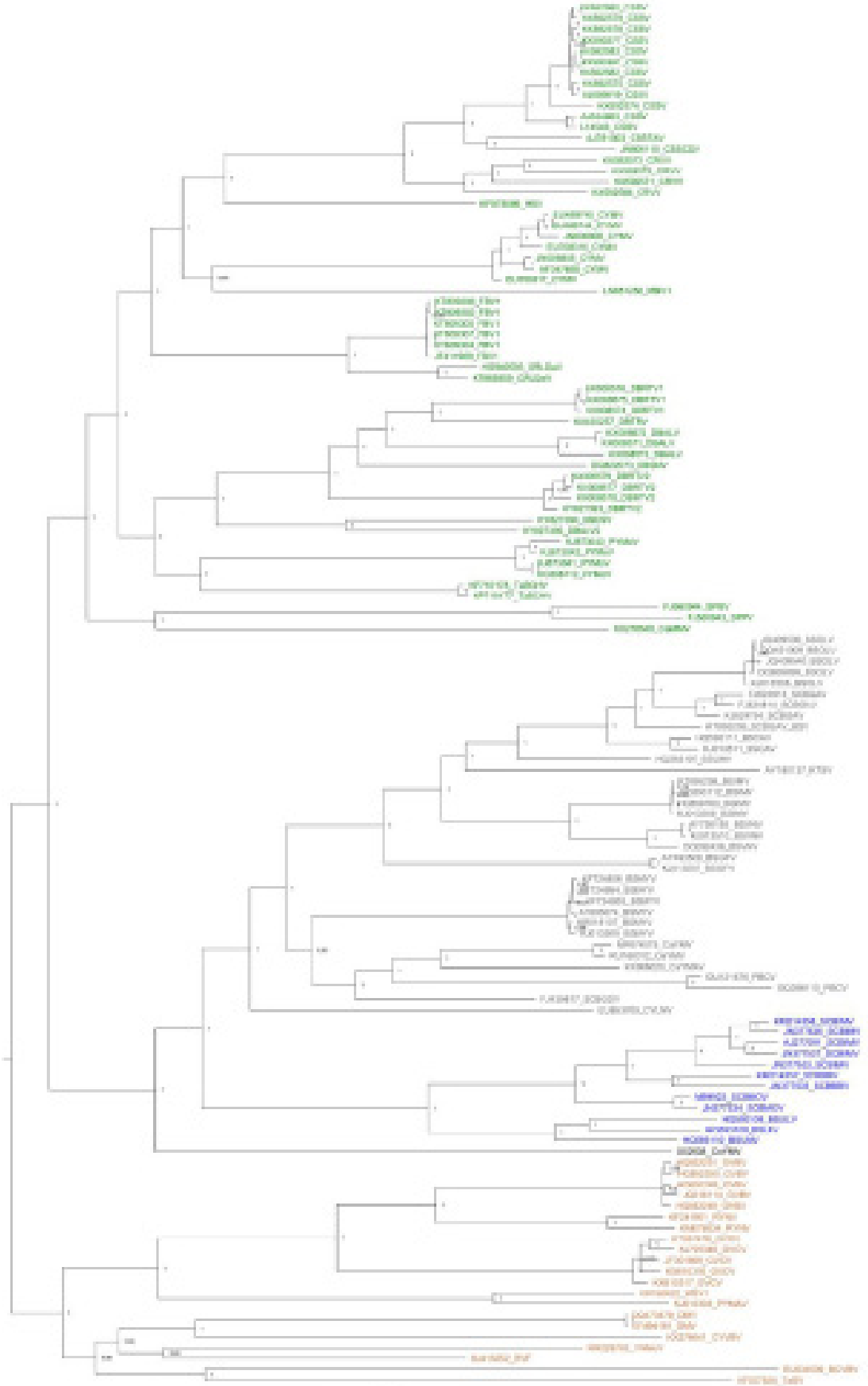


Figure 3. Non-tree-like evolution based on the complete genome nucleotide sequences of badnaviruses. Badnavirus groups 1 (indicated in grey), 2 (green), 3 (blue) and 4 (orange). Commelina yellow mottle virus (CoYMV; highlighted in grey) was placed apart from the four main groups.

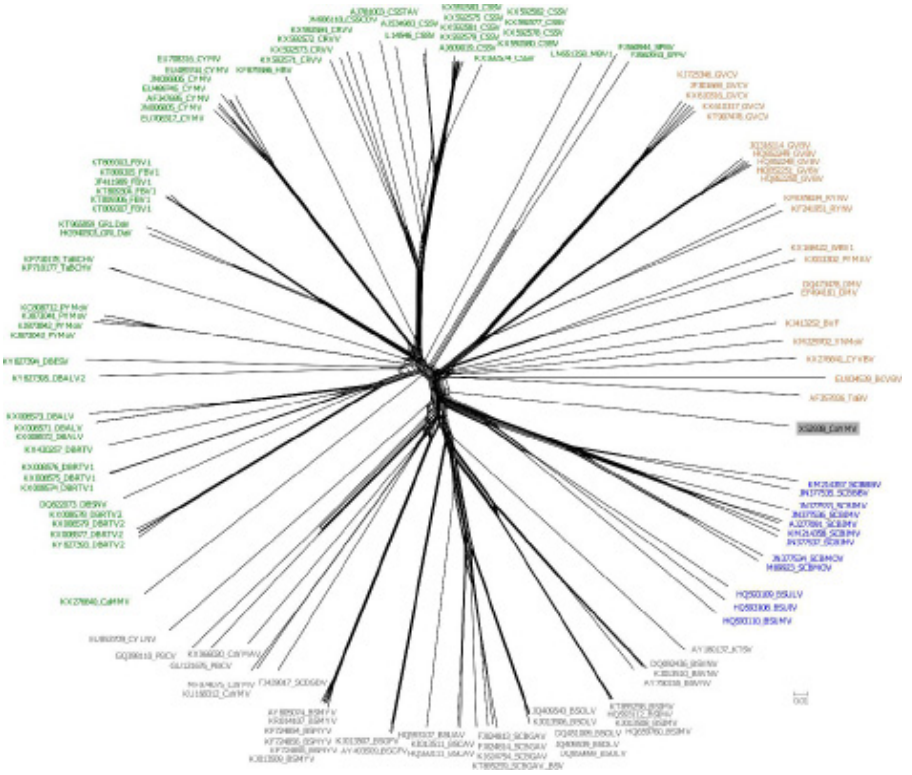
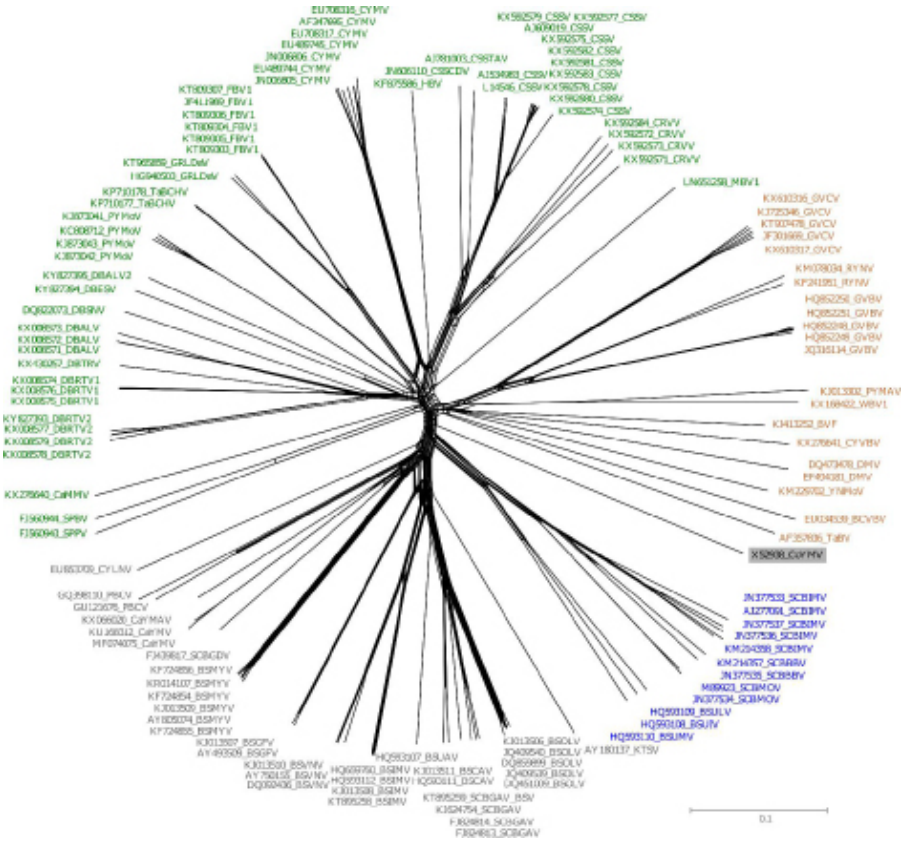


Figure 4. Non-tree-like evolution based on the ORFIII nucleotide sequences of badnaviruses. Badnavirus groups 1 (indicated in grey), 2 (green), 3 (blue) and 4 (orange). Commelina yellow mottle virus (CoYMV; highlighted in grey) was placed apart from the four main groups.



To investigate putative parental sequences and recombination breakpoints, the complete genome data set was analyzed using the RDP4 package. Based on a stringent set of criteria, at least 23 independent recombination events were detected among badnavirus isolates, with ten of them involving species associated with sugarcane and banana (Table 2). Additional recombination events involved viruses infecting citrus, cacao, black pepper, pineapple and grapevine. Most recombination breakpoints were located in ORF III and in the intergenic region (Table 2).

Table 2. Putative recombination events detected within badnavirus isolates, based on complete genome sequences.

Event	Breakpoints*		Recombinant	Parents		Methods†	*P* value‡
	Begin	End		Minor	Major		
1	202	1499	JN006805_CYMV	JN006806_CYMV	AF347695_CYMV	RGBMCS3	1.7 E-82
2	833	1183	AY805074_BSMYV	Unknown	KR014107_BSMYV	RGBMCS3	8.6 E-20
3	3493	7233	^KJ873042_PYMoV	KJ873043_PYMoV	KC808712_PYMoV	RGBMCS3	6.4 E-28
4	7283(?)	202(?)	^EU489745_CYMV EU489744_CYMV	JN006805_CYMV	JN006806_CYMV	RGBMCS3	7.3 E-20
5	202(?)	4651	^JN006806_CYMV	EU489745_CYMV	AF347695_CYMV	RGBMCS3	1.7 E-44
6	7510	80	HQ852251_GVBV	Unknown	HQ852249_GVBV	RGBMCS3	5.3 E-15
7	4342	4833(?)	^HQ852250_GVBV HQ852248_GVBV	Unknown	HQ852249_GVBV	RGBMCS3	3.1 E-14
8	348	499	^KJ013508_BSIMV	Unknown	KT895258_BSIMV	RGBMCS	3.1 E-13
9	1007	1186	^JN377534_SCBMOV	KM214358_SCBIMV	M89923_SCBMOV	RGBMCS3	4.6 E-12
10	3733	3848	^JN377533_SCBIMV	Unknown	KM214357_SCBBBV	RGBMCS	7.1 E-11
11	1066(?)	1309	^GU121676_PBCV	Unknown	GQ398110_PBCV	RBMC3	5.0 E-06
12	6306	647(?)	^KF724854_BSMYV KF724856_BSMYV KJ013509_BSMYV	Unknown	AY805074_BSMYV	RGMC3	1.5 E-10
13	7050	7146(?)	^JN377534_SCBMOV AJ277091_SCBIMV M89923_SCBMOV KM214357_SCBBBV	FJ824813_SCBGAV	Unknown	RGBS3	1.8 E-10
14	4356	4863	^DQ092436_BSVNV AY750155_BSVNV KJ013510_BSVNV	Unknown	KJ013508_BSIMV	RBMC3	2.1 E-10
15	865	1234	HQ593107_BSUAV	HQ593109_BSULV	HQ593111_BSCAV	RBMC3	2.6 E-10
16	7463(?)	202(?)	^AF347695_CYMV JN006805_CYMV	Unknown	EU708317_CYMV	RGBMC3	2.1 E-09
17	7290(?)	7401	^DQ092436_BSVNV AY750155_BSVNV	FJ439817_SCBGDV	HQ659760_BSIMV	RGCS3	2.0 E-08
18	7078	210	^AJ534983_CSSV L14546_CSSV	Unknown	KX592583_CSSV KX592582_CSSV	RBMC3	8.8 E-08
19	1400(?)	2363	AY493509_BSGFV KJ013507_BSGFV	KU168312_CaYMV	AY180137_KTSV	RBMCS	2.2 E-09
20	5022	6244(?)	^KT965859_GRLDaV HG940503_GRLDaV	FJ824813_SCBGAV	KX592572_CRVV	BMCS3	4.9 E-07
21	601	1206	^JN377536_SCBIMV KM214358_SCBIMV AJ277091_SCBIMV JN377537_SCBIMV	M89923_SCBMOV	KM214357_SCBBBV	RGBMCS	2.6 E-10
22	393	702	^HQ593107_BSUAV	Unknown	DQ092436_BSVNV	RBMCS3	4.7 E-05
23	1055	1388(?)	^KX610317_GVCV	KJ725346_GVCV	JF301669_GVCV	RBMCS	2.4 E-06

* Numbering starts at the 5' end of the minus-strand primer-binding site and increases clockwise. (?), breakpoint could not be precisely pinpointed.

† R, RDP; G, GeneConv; B, Bootscan; M, MaxChi; C, Chimera; S, SisScan; 3, 3SEQ.

‡ The reported *P* values are for the methods indicated in red, and they are the lowest *P* values calculated for the region in question.

^ The recombinant sequence may have been misidentified (one of the identified parents might be the recombinant).

Population genetic variability

A total of 299 RT/RNaseH partial sequences were obtained for 11 distinct badnavirus species, and the nucleotide diversity of each dataset/population was calculated (Table 3). Additionally, 96 *Dioscorea bacilliform AL virus* (DBALV) partial sequences from Guimarães et al. (2015) were added.

Table 3. Genetic variability of the RT/RNaseH region of badnavirus populations infecting distinct hosts.

Population	No. of sequences	Lenght (nt)	π* (SD)†
BSGFV	13	668	0.0209 (±0.0069)
BSMYV	10	659	0.0054 (±0.0013)
BSOLV	35	668	0.0164 (±0.0038)
BSUAV	8	662	0.0456 (±0.0039)
BSULV	12	668	0.1467 (±0.0133)
BSUMV	8	668	0.0579 (±0.0061)
PYMoV	27	579	0.0944 (±0.0037)
SPBVB	35	668	0.0418 (±0.0088)
CSSV	26	575	0.0334 (±0.0055)
CSSCDV	7	575	0.0162 (±0.0019)
SCBIMV	118	725	0.0876 (±0.0034)
DBALV‡	96	435	0.0534 (±0.0054)

* Pairwise, per-site nucleotide diversity.

† Standard deviation.

‡ Guimarães et al., 2015.

High values of per-site nucleotide diversity were observed for *Banana streak UA virus* (BSUAV), *Banana streak UL virus* (BSULV), *Banana streak UM virus* (BSUMV), *Piper yellow mottle virus* (PYMoV), SPBV and SCBIMV, suggesting high genetic variability for the partial RT/RNaseH region in different badnavirus species/populations (Table 3). The most variable sequence data sets were BSULV (π = 0.1467), PYMoV (π = 0.0944) and SCBIMV (π = 0.0876), being higher than the previously reported DBALV (π = 0.0534).

DISCUSSION

Badnaviruses have been reported infecting economically important crops worldwide (Eni et al., 2008; James et al., 2011; Johnson et al., 2012; Kouakou et al., 2012; Deeshma and Bhat, 2015; Silva et al., 2015), causing great losses in cacao (Kouakou et al., 2012) and banana (Dahal et al., 2000; Daniells et al., 2001). They comprise a highly diversified group, and recent studies have shown that badnavirus populations exhibit high levels of genetic variability (Muller et al., 2011; Karuppaiah et al., 2013; Guimarães et al., 2015; Silva et al., 2015). Nucleotide sequence comparisons of the partial RT/RNaseH domains (579pb) and a threshold of ≥80% identity have largely been used for species demarcation in the genus *Badnavirus* (Muller et al., 2011; Geering and Hull, 2012; Guimarães et al., 2015; Silva et al., 2015), with universal primers described by Yang et al., (2003) being very efficient to amplify this genomic region. However, the existence of closely related badnavirus species which share more than 80% nucleotide identity in that region suggests this criterion needs to be revised (Harper et al., 2005; Muller et al., 2011; Karuppaiah et al., 2013; Silva et al., 2015).

In the present study, nucleotide sequence comparisons of a large sequence data set showed that the threshold of ≥80% identity for the RT/RNaseH genomic region [partial (579pb) or full (1020pb)] allowed the species demarcation of most badnavirus isolates with full-length genome sequences available in GenBank. However, as previously reported, it was not possible to distinguish a few badnaviruses from banana and sugarcane (Muller et al., 2011; Karuppaiah et al., 2013; Silva et al., 2015), even when analyzed the entire ORF III (which corresponds to ~80% of the badnavirus genome).

We suggest a few alternatives to solve the identification problems observed for badnaviruses infecting banana and sugarcane. First, maintaining the criterion of ≥80% nucleotide sequence identity for the RT/RNaseH genomic region, closely related banana streak viruses (BSV) and sugarcane bacilliform

viruses (SCBV) species sharing ≥80% nucleotide sequence identity should be considered as different strains of a same species. Tanking account biological aspects as host range (Geering and Hull, 2012), cross infection involving sugarcane and banana-infecting badnaviruses has been observed (Lockhart and Autrey, 1988; Bouhida et al., 1993; Jones and Lockhart, 1993), which reinforces the ideia SCBV and BSV sharing more than 80% identity may belong to a same viral species. Additionally, the isolates SPPV (#access FJ560943) and SPBVB (#access FJ560944), previously reported as distinct species, must belong to a same badnavirus species as they share more than 80% identity for RT/RNaseH and the same host range, besides their narrow phylogenetic relationship. Second, a new threshold for species demarcation based on nucleotide identity of the RT/RNaseH domain could be established. In our analysis, it was observed the badnavirus species showed no more than 82,5% nucleotide identity for the RT/RNaseH domain, and therefore higher values could be adopted as the new threshold for species demarcation. Third, considering the inefficiency to differentiate some badnavirus species using the currently established criteria, full-length genome sequences could be used for taxonomy.

Phylogenetic analysis showed clear shuffle of SCBVs and BSVs isolates into at least two distinct badnavirus clusters (*i.e.*, badnavirus groups 1 and 3 *sensu* Muller et al., 2011). The close genetic relationship among sugarcane and banana-infecting badnaviruses, and the polyphyletic structure of these viruses (Gayral and Iskra-Caruana, 2009; Muller et al., 2011), strongly support the hypothesis of a host shift, although it is not possible to determine whether sugarcane or banana was the original host (Gayral and Iskra-Caruana, 2009; Muller et al., 2011).

The phylogenetic relationships observed for all data sets analyzed, in which SCBVs and BSVs isolates are grouped in sister clusters, agree with results of pairwise sequence comparisons. Besides the three well defined badnavirus groups reported by Muller et al., (2011), a fourth phylogenetic cluster is proposed here, named badnavirus group 4. In the phylogeny based on full RT/RNaseH, the groups 1 and 3 (both composed predominantly by sugarcane and banana-infecting viruses) are more distantly related, with group 3 being closer to group 4. However, in phylogenies based either on ORFIII or complete genomes, a close relationship between groups 1 and 3 was evident, so they can be considered as sister groups. These results reinforce the hypothesis that sugarcane and banana-infecting badnaviruses have a common evolutionary history.

The high levels of nucleotide diversity for the partial RT/RNaseH region in some badnavirus species (BSUAV, BSULV, BSUMV, PYMoV, SPBVB, SCBIMV and DBALV) are comparable to values observed for RNA viruses (Guimarães et al., 2015). However, this variability is lower when compared to less conserved regions [ORFs I, II and intergenic region (IGR)], with IGR showing the highest values of genetic diversity (0.468) followed by ORFII 0.367; Sharma et al., 2015). These results suggest although the partial RT/RNaseH region be variable in different badnavirus populations, non-conserved regions tend to have higher nucleotide diversity.

The high genetic variability observed for badnaviruses has been attributed to error-prone replication by their reverse transcriptase (Bousalem et al., 2008). Reverse transcriptases (RT) are known to produce errors in retroviruses and retroelements for which the fidelity rates have been estimated (Svarovskaia et al., 2003). Although fidelity rates of caulimovirid RTs have not been estimated, it is believed that the lack of proofreading activity contributed to the high mutation rates observed for these viruses (Svarovskaia et al., 2003). Nevertheless, the contribution of recombination to the genetic diversity of badnaviruses must also be considered (Govind et al., 2014).

One of the basic assumptions for successful recombination is the occurrence of mixed infections, with the presence of the viruses in the same host cell (Zhou et al., 1997; García-Andrés et al., 2006; Graham et al., 2010). Cross infection of badnaviruses seems to be an uncommon event, however it has been reported for viruses infecting sugarcane and banana, which could explain the putative recombinant origin of SCBV and BSV isolates (Lockhart and Autrey, 1988; Bouhida et al., 1993; Sharma et al., 2015; Bath et al., 2016). Here, besides the putative recombination events affecting the evolution of SCBV and BSV isolates, it also seems an important evolutionay mechanism for diversification of cacao-infecting badnaviruses.

Nucleotide sequence comparisons of the partial RT/RNaseH region are sufficient for species demarcation of most badnaviruses currently known. However, the ≥80% nucleotide identity threshold alone

is unable to differentiate all SCBV and BSV species and should be reviewed. Finally, mutation and putative recombination events may be involved with the high levels of genetic variability and diversity observed in *Badnavirus*.

ACKNOWLEDGMENTS

C.H.L.H.F. was the recipient of a CAPES master's degree fellowship; the authors wish to thank F.M. Zerbini for critical review of the manuscript; this work was supported by FAPEAL.

REFERENCES

Bhat, A.I.; Hohn, T.; Selvarajan, R. Badnaviruses: The Current Global Scenario. *Viruses,* **2016**, 177, 1-29.

Bouhida, M.L.; Lockhart, B.E.; Olszewski, N.E. An analysis of the complete sequence of a sugarcane bacilliform virus genome infectious to banana and rice. *Journal of General Virology*, **1993**, 74, 15-22.

Bousalem, M.; Douzery, E.J.P.; Seal, S.E. Molecular taxonomy, phylogeny, and evolution of plant reverse transcribing viruses (Caulimoviridae) inferred from the reverse transcriptase sequences. *Archives of Virology*, **2008**, 153, 1085-1102.

Cheng, C.P.; Lockhart, B.E.; Olszewski, N.E. The ORF I and II Proteins of Commelina yellow mottle virus are virion-associated. *Virology*, **1996**, 223, 263-271.

Côte, F.X.; Galzi, S.; Folliot, M.; Lamagnère, Y.; Teycheney, P.Y.; Iskra-Caruana, M.L. Micropropagation by tissue culture triggers differential expression of infectious endogenous Banana streak virus sequences (eBSV) present in the B genome of natural and synthetic interspecific banana plantains. Molecular Plant *Pathology*, **2010**, 11, 137-144.

Dahal, G.; Ortiz, R.; Tenkouano, A.; Hughes, Jd'A.; Thottappilly, G.; Vuylsteke, D.; Lockhart, B.E.L. Relationship between natural occurrence of banana streak badnavirus and symptom expression, relative concentration of viral antigen, and yield characteristics of some micropropagated *Musa* spp. *Plant Pathology,* **2000**, 49, 68-79.

Dallot, S.; Accuna, P.; Rivera, C.; Ramirez, P.; Cote, F.; Lockhart, B.E.L.; Caruana, M.L. Evidence that the proliferation stage of micropropagation procedure is determinant in the expression of Banana streak virus integrated into the genome of the FHIA21 hybrid (Musa AAAB). *Archives of Virology*, **2001**, 146, 2179-2190.

Daniells, J.W.; Geering, A.D.W.; Bride, N.J.; Thomas, J.E. The effect of banana streak virus on the growth andyield of dessert bananas in tropical Australia. *Annals of Applied Biology*, **2001**, 139, 51-60.

Deeshma, K.P.; Bhat, A.I. Complete genome sequencing of Piper yellow mottle virus infecting black pepper, betelvine, and Indian long pepper. *Virus Genes*, **2015**, 50, 172–175.

Edgar R.C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics,* **2004**, 5, 1–19.

Eni, A.O.; Hughes, Jd.; Asiedu, R.; Rey, M.E. Sequence diversity among Badnavirus isolates infecting yam (Dioscorea spp.) in Ghana, Togo, Benin and Nigeria. *Archives of Virology*, **2008**, 153, 2263-2272.

García-Andrés, S.; Monci, F.; Navas-Castillo, J.; Moriones, E. Begomovirus genetic diversity in the native plant reservoir Solanum nigrum: Evidence for the presence of a new virus species of recombinant nature. *Virology*, **2006**, 350, 433-442.

Gayral, P.; Iskra-Caruana, M. Phylogeny of Banana Streak Virus Reveals Recent and Repetitive Endogenization in the Genome of Its Banana Host (Musa sp.). *Journal of Molecular Evolution*, **2009**, 69, 65-80.

Graham, A.P.; Martin, D.P.; Roye, M.E. Molecular characterization and phylogeny of two begomoviruses infecting Malvastrum americanum in Jamaica: Evidence of the contribution of inter-species recombination to the evolution of malvaceous weed-associated begomoviruses from the Northern Carribbean. *Virus Genes*, **2010**, 40, 256-266.

Geering, A.D.W.; Hull, R. Family Caulimoviridae. In: King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ (Eds.). Virus Taxonomy. 9th Report of the International

Committee on Taxonomy of Viruses. London, UK. Elsevier Academic Press. **2012**, pp. 429-443.

Geering, A.D.W.; Mcmichael, L.A.; Dietzgen, R.G.; Thomas, J.E. Genetic diversity among Banana streak virus isolates from Australia. *Phytopathology*, **2000**, 90, 921-927.

Geijskes, R.J.; Braithwaite, K.S.; Dale, J.L.; Harding, R.M.; Smith, G.R. Sequence analysis of an Australian isolate of sugarcane bacilliform badnavirus. *Archives of Virology*, **2002**, 147, 2393-2404.

Govind, P.R.; Susheel, K.S.; Deepti, S.; Meenakshi, A.; Priyanka, S.; Virendra, K.B. Genetically Diverse Variants of Sugarcane bacilliform virus Infecting Sugarcane in India and Evidence of a Novel Recombinant Badnavirus Variant. *Journal of Phytopathology*, **2014**, 162, 779–787.

Guimarães, K.M.; Silva, S.J.C.; Melo, A.M.; Ramos-Sobrinho, R.; Lima, J.S.; Zerbini, F.M.; Assunção, I.P.; Lima, G.S.A. Genetic variability of badnaviruses infecting yam (Dioscorea spp.) in northeastern Brazil. *Tropical Plant Pathology*, **2015**, 40, 111-118.

Hagen, L.S.; Jacquemond, M.; Lepingle, A.; Lot, H.; Tepfer, M. Nucleotide sequence and genomic organisation of cacao swollen shoot virus. *Virology*, **1993**, 196, 619-628.

Harper, G.; Hart, D.; Moult, S.; Hull, R.; Geering, A.; Thomas, J. Diversity of banana streak virus isolates in Uganda. *Archives of Virology*, **2005**, 150, 2407-2420.

Harper, G.; Hull, R. Cloning and sequence analysis of Banana Streak Virus DNA. *Virus Genes*, **1998**, 17, 271-278.

Hohn, T.; Richert-Poggeler, K.R.; Harper, G.; Schawarzacher, T.; Teo, C.; Teycheney, P.Y.; Iskra-Caruana, M.L.; Hull, R. Evolution of integrated plant viruses. In: Roosinck M (Eds.) Plant Virus Evolution. Heidelberg, Germany. Academic Springer. **2008**, pp. 54–76.

Huson, D.H.; Bryant, D. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, **2006**, 23, 254–267.

Jacquot, E.; Hagen, L.S.; Jacquemond, M.; Yot, P. The open reading frame 2 product of cacao swollen shoot badnavirus is a nucleic acid-binding protein. *Virology*, **1996**, 225, 191-195.

James, A.P.; Geijskes, R.J.; Dale, J.L.; Harding, R.M. Molecular characterization of six badnavirus species associated with leaf streak disease of banana in East Africa. *Annals of Applied Biology*, **2011**, 158, 346-353.

Johnson, A.M.; Borah, B.K.; Sai Gopal, D.V.; Dasgupta, I. Analysis of full-length sequences of two Citrus yellow mosaic badnavirus isolates infecting Citrus jambhiri (Rough Lemon) and Citrus sinensis L. Osbeck (Sweet Orange) from a nursery in India. *Virus Genes*, **2012**, 45, 600–605.

Jones, D.R.; Lockhart, B.E.L. Banana streak disease. Musa fact sheet No. 1, International network for improvement of banana and plantain, Montpellier, France. **1993**.

Karuppaiah, R.; Viswanathan, R.; Kumar, V.G. Genetic diversity of Sugarcane bacilliform virus isolates infecting Saccharum spp. in India. *Virus Genes*, **2013**, 46, 505–516.

Kouakou, K.; Kébé, B.I.; Kouassi, N.; Muller, E. Geographical distribution of Cacao swollen shoot virus molecular variability in Côte d'Ivoire. *Plant Disease*, **2012**, 96, 1445-1450.

Kumar, S.; Stecher, G.; Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for Bigger Datasets. *Molecular Biology and Evolution*, **2016**, 33, 1870-1874.

Lheureux, F.; Carreel, F.; Jenny, C.; Lockhart, B.E.L.; Iskra-Caruana, M.L. Identification of genetic markers linked to banana streak disease expression in inter-specific Musa hybrids. *Theoretical and Applied Genetics*, **2003**, 106, 594-598.

Lockhart, B.E.L.; Autrey, L.J.C. Occurrence in sugarcane of a bacilliform virus related serologically to banana streak virus. *Plant Disease*, **1988**, 72, 230-233.

Martin, D.P.; Murrell, B.; Golden, M.; Khoosal, A.; Muhire, B. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evolution*, **2015**, 1, 1-5.

Medberry, S.L.; Lockhart, B.E.L.; Olszewski, N.E. Properties of Commelina yellow mottle virus complete DNA sequence, genomic discontinuities and transcript suggest that it is a pararetrovirus. *Nucleic Acids Research*, **1990**, 18, 5505-5513.

Miller, M.A.; Holder, M.T.; Vos, R.; Midford, P.E.; Liebowitz, T.; Chan, L.; Hoover, P.; Warnow, T. (**2010**) The CIPRES portals. CIPRES. http://www.phylo.org/sub-sections/portal (Accessed 15 Aug 2016).

Muller, E.; Dupuy, V.; Blondin, L.; Bauffe, F.; Daugrois, J.H.; Laboureau, N.; Iskra Caruana, M.L. High molecular variability of sugarcane bacilliform viruses in Guadeloupe implying the existence of at least three new species. *Virus Reseach*, **2011**, 160, 414-419.

Muhire, B.; Martin, D.P.; Brown, J.K.; Navas-Castillo, J.; Moriones, E.; Zerbini, F.M.; Rivera-Bustamante, R.; Malathi, V.G.; Briddon, R.W.; Varsani, A. A genome-wide pairwise-identity-based proposal for the classification of viruses in the genus Mastrevirus (family Geminiviridae). *Archives of Virology*, **2013**, 158, 1411-1424.

Rannala, B.; Yang, Z. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *Journal of Molecular Evolution*, **1996**, 43, 304-311.

Ronquist, F.; Teslenko, M.; van der Mark, P.; Ayres, D.L.; Darling, A.; Hohna, S.; Larget, B.; Liu, L.; Suchard, M.A.; Huelsenbeck, J.P. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice across a Large Model Space. *Systematic Biology Advance Access,* **2012**, 61, 539-42.

Rozas, J.; Ferrer-Mata, A.; Sánchez-DelBarrio, J.C.; Guirao-Rico, S.; Librado, P.; Ramos-Orsins, S.E.; Sánchez-Gracia, A. DnaSP v6: DNA Sequence Polymorphism Analysis of Large Datasets. *Molecular Biology and Evolution,* **2017**, 34, 3299-3302.

Sharma, S.K.; Kumar, P.V.; Geetanjali, A.S.; Pun, K.B.; Baranwal, V.K. Subpopulation level variation of banana streak viruses in India and common evolution of banana and sugarcane badnaviruses. *Virus genes*, **2015**, 50, 450-465.

Silva, J.M.; Jobim, L.J.; Ramos-Sobrinho, R.; Lima, J.S.; Assunção, I.P.; Cruz, M.M.; Lima, G.S.A. Incidence and species diversity of badnaviruses infecting sugarcane from a germplasm collection in Brazil. *Tropical Plant Pathology*, **2015**, 40, 212-217.

Staginnus, C.; Iskra-Caruana, M.; Lockhart, B.; Hohn, T.; Richert-Pöggeler, K.R. Suggestions for a nomenclature of endogenous pararetroviral sequences in plants. *Archives of Virology*, **2009**, 154, 1189-1193.

Svarovskaia, E.S.; Cheslock, S.R.; Zhang, W.H.; Hu, W.S.; Pathak, V.K. Retroviral mutation rates and reverse transcriptase fidelity. Fronties in Bioscience. 2003, 8, 117–134.

Yang, I.C.; Hafner, G.J.; Dale, J.L.; Harding, R.M. Genomic characterization of taro bacilliform virus. *Archives of Virology*, **2003**, 148, 937-949.

Zhou, X.; Liu, Y.; Calvert, L.; Munoz, C.; Otim-Nape, G.W.; Robinson, D.J.; Harrison, B.D. Evidence that DNA-A of a geminivirus associated with severe cassava mosaic disease in Uganda has arisen by interspecific recombination. *Journal of General Virology*, **1997**, 78, 2101-2111.