J Mol Evol (2009) 69:65–80
DOI 10.1007/s00239-009-9253-2

Phylogeny of *Banana Streak Virus* Reveals Recent and Repetitive Endogenization in the Genome of Its Banana Host (*Musa* sp.)

Philippe Gayral · Marie-Line Iskra-Caruana

Received: 6 February 2009 / Accepted: 26 May 2009 / Published online: 11 June 2009

Abstract *Banana streak virus* (BSV) is a plant dsDNA pararetrovirus (family *Caulimoviridae*, genus *badnavirus*). Although integration is not an essential step in the BSV replication cycle, the nuclear genome of banana (*Musa* sp.) contains BSV endogenous pararetrovirus sequences (BSV EPRVs). Some BSV EPRVs are infectious by reconstituting a functional viral genome. Recent studies revealed a large molecular diversity of episomal BSV viruses (i.e., nonintegrated) while others focused on BSV EPRV sequences only. In this study, the evolutionary history of *badnavirus* integration in banana was inferred from phylogenetic relationships between BSV and BSV EPRVs. The relative evolution rates and selective pressures (d_N/d_S ratio) were also compared between endogenous and episomal viral sequences. At least 27 recent independent integration events occurred after the divergence of three banana species, indicating that viral integration is a recent and frequent phenomenon. Relaxation of selective pressure on badnaviral sequences that experienced neutral evolution after integration in the plant genome was recorded. Additionally, a significant decrease (35%) in the EPRV evolution rate was observed compared to BSV, reflecting the difference in the evolution rate between episomal dsDNA viruses and plant genome. The comparison of our results with the evolution rate of the *Musa* genome and other reverse-transcribing viruses suggests that EPRVs play an active role in episomal BSV diversity and evolution.

P. Gayral · M.-L. Iskra-Caruana (✉)
CIRAD, UMR Biologie et Génétique des Interactions Plante-Parasite (BGPI), TA A-54/K, 34398 Montpellier Cedex 5, France
e-mail: marie-line.caruana@cirad.fr

Keywords *Badnavirus* · Banana (*Musa* sp.) · *Banana streak virus* (BSV) · d_N/d_S ratio · Endogenous pararetrovirus (EPRV) · Evolution rate · Integration · Selective constraints

Introduction

Recent studies revealed that endogenous viral sequences are widespread in the nuclear genome of plants. They originate from ancient viral infections that become fixed in the germline. The mechanism of integration is still unknown but may involve illegitimate recombination, as no plant viruses described up to now require an integration step for replication (Staginnus and Richert-Poggeler 2006). Endogenous pararetrovirus sequences (EPRVs) are the most abundant type of endogenous viral sequences in plants. They have been found in the nuclear genome of distantly related plant families, such as bitter orange (*Poncirus trifoliata*), potato and relatives (*Solanum* sp.), petunia (*Petunia* sp.), tobacco and relatives, rice (*Oryza sativa*), and banana (*Musa* sp.) (reviewed in Staginnus and Richert-Poggeler 2006), and more recently in lucky bamboo (*Dracaena sanderiana*) (Su et al. 2007) and Dahlia (*Dahlia variabilis*) (Pahalawatta et al. 2008). EPRVs are related to the *Caulimoviridae* family, also called plant pararetroviruses (PRV), which have a circular double stranded DNA (dsDNA) genome (7.0–8 kbp). This viral family is composed of six genera, and EPRVs have been described in most of them: *Petuvirus*, *Cavemovirus*, *Badnavirus*, *Tungrovirus*, and *Caulimovirus*.

Because the number of EPRV copies in tobacco was in the thousands, they were described as a novel class of dispersed repetitive elements that have a significant impact on the complexity and evolution of the host

genomes (Hohn et al. 2008; Jakowitsch et al. 1999). All EPRVs described have a similar rearranged pattern with tandem repeats, internal duplications, fragmentations, and inversions of viral genomes (Gayral et al. 2008; Ndowora et al. 1999; Richert-Poggeler et al. 2003). The majority of EPRVs result in partial and nonfunctional viral genomes (Geering et al. 2005a; Kunii et al. 2004). However, several integrations contain a full-length viral genome with functional open reading frames (ORFs). Such uncorrupted EPRV sequences can then be activated, resulting in the release of functional viral genomes that infect the host plant. Infectious EPRVs have been reported for *Petunia vein clearing virus* (PVCV) in Petunia—*Petunia hybrida* (Richert-Poggeler and Shepherd 1997), for *Tobacco vein clearing virus* (TVCV) in the hybrid tobacco species *Nicotiana edwardsonii* (Lockhart et al. 2000), and for *Banana streak virus* (BSV) in banana—*Musa* sp. (Ndowora et al. 1999). Two main mechanisms are proposed to explain EPRV activation: homologous recombination between repeat regions surrounding EPRVs and causing the excision of a circular viral genome (Gaut et al. 2007; Ndowora et al. 1999; Schuermann et al. 2005) and direct transcription of EPRVs leading to a functional viral pregenomic RNA (Noreen et al. 2007; Richert-Poggeler et al. 2003). Furthermore, because EPRV transcription and subsequent siRNA occurs in *Petunia* sp., *Nicotiana* sp., and *Solanum* sp. (Hansen et al. 2005; Mette et al. 2002; Noreen et al. 2007; Staginnus et al. 2007), it has been suggested that EPRVs could play a role in plant resistance to the cognate virus via homology-dependant gene silencing (Hull et al. 2000; Mette et al. 2002; Staginnus and Richert-Poggeler 2006). However, direct experimental evidence is required to confirm this hypothesis.

BSV are nonenveloped bacilliform viruses that cause banana streak disease in all banana-producing areas (Lockhart and Jones 2000). This pathogen appears to be a very useful model to study viral evolution and the consequences of integration that cause rapid changes in selective pressures. Indeed, BSV exists not only as an episomal pathogen species transmitted horizontally by mealybugs, which infects wild and domesticated banana species, but also as EPRV sequences in the *Musa* genome. BSV EPRVs were found in the genome of several cultivated banana genotypes originating from three closely related *Musa* species: *M. acuminata* (A genome), *M. balbisiana* (B genome), and *M. schizocarpa* (S genome) (Geering et al. 2001, 2005a), all belonging to the *Eumusa* section (Ude et al. 2002). Studies of episomal virus sequences during epidemics in Uganda (Harper et al. 2004, 2005), Australia (Geering et al. 2000), and Mauritius (Jaufeerally-Fakim et al. 2006) revealed great genetic diversity among BSV. Based on sequence analysis of a partial reverse transcriptase (RT)/RNase H gene located in ORF3 (580 bp), obtained by PCR using degenerate primers, the 'Caulimoviridae study group' of the International Committee on Taxonomy of Viruses (ICTV) defined a threshold of 20% nucleotide diversity to differentiate two episomal badnavirus species (Fauquet et al. 2005). Based on this definition, BSV is composed of several distinct virus species able to infect the same *Musa* host plant, rather than of several strains of the same viral species. Furthermore, no significant genetic exchanges were detected between BSV species (E. Muller, pers. comm.), suggesting the presence of a species barrier. At the same time, Geering et al. (2005a) searched for badnavirus sequences integrated in the *Musa* genome using a degenerate PCR approach with virus-free plant material. These authors also observed a high diversity of endogenous BSV-related sequences in several *Musa* species. Nevertheless, no phylogenetic studies have been performed so far to investigate the relationships between episomal virus particles and EPRV sequences. This step is critical to understanding the evolutionary history of badnaviruses, which are considered as emerging pathogens in tropical countries, and the phenomenon of viral integration in plants. In this study, the terms BSV EPRVs and badnavirus EPRVs refer to endogenous viral sequences, whereas BSV, badnavirus, and PRV (pararetrovirus) refer to episomal viruses.

This work first aimed to infer a robust phylogeny from all published episomal and EPRV sequence: 13 species of badnavirus, 99 EPRV sequences from several *Musa* species and genotypes, and 105 BSV sequences of uncertain origin. Co-diversification between BSV and banana was also assessed by comparing badnavirus phylogeny and *Musa* speciation events. Importantly, direct comparison between endogenous and episomal viral sequences was made possible by the use of a single phylogeny for the two types of viral sequences types. We then used this phylogenetic framework to address several evolutionary issues concerning EPRVs. What is the distribution of endogenous viral sequences: are they present throughout the genetic diversity of BSV or are they clustered in a small number of BSV groups? Are integration events rare or frequent (i.e., how many independent integration events have occurred in the *Musa* genome)? Are these integration events ancient (i.e., are they shared between *Musa* species), or recent (i.e., specific to one *Musa* species or genotype)? Finally, we studied the effects of integration events on viral DNA evolution by comparing the evolution rate and selective pressure acting on episomal sequences—evolving in the context of host–parasite interactions vs. EPRVs—that are part of the host genome.

Materials and Methods

Nucleotide Sequences

A 540-bp fragment of the RT/RNase H region located in ORF3 in the genome of badnaviruses was used in this study (Table 1). The sequences corresponded either to episomal viruses and were labeled 'PRV' or to endogenous sequences and were labeled 'EPRV.' The 'PRV' category was composed of sequences retrieved from full length published viral genomes of both BSV and six closely related badnaviruses species. In this study, we generated one additional sequence of the episomal *Banana streak cavendish virus* (BSCavV) (GenBank accession numbers shown in Table 1). This sequence was generated from a BSCavV-infected *Musa acuminata* cv. Williams after a multiplex-immuno-capture-PCR (Le Provost et al. 2006) that amplified episomal viral particles only with primers BadnaFP 5′-GCCITTYGGIITIAARAAYGCICC-3′ and BadnaRP 5′-CCAYTTRCAIACISCICCCCAICC-3′ (Yang et al. 2003) at annealing temperature (T_a) of 55°C. The product was cloned and 13 positive clones were sequenced using M13F universal primer. The sequences showed more than 98% similarity. A single sequence was chosen for this study.

Sequences of the 'EPRV' category are endogenous sequences from both BSV and badnaviruses. A total of 87 EPRV sequences retrieved from GenBank were originally generated from a degenerate PCR approach using total DNA of virus-free *Musa acuminata*, *Musa balbisiana*, and *Musa schizocarpa* accessions (Geering et al. 2005a). Eleven additional EPRV sequences were generated in this study from *Musa acuminata* and *Musa balbisiana* checked for BSV-free status (data not shown) by multiplex-immuno-capture-PCR (Le Provost et al. 2006). PCR amplification with total DNA of *Musa balbisiana* cv. Pisang Klutuk Wulung (PKW) was performed using primers Badna1 5′-CTNTAYGARTGGYTNGTNATGCCNTTYG G-3′ and GfR 5′-TCGGTGGAATAGTCCTGAGTCTTC-3′ at $T_a = 51$°C and with 25 pmol (instead of 10 pmol) of primer Badna1 in the PCR mix, and the product was cloned. Two clones (PKW514 and PKW515) were sequenced in both orientations with the universal primers M13F and M13R (GenBank accession numbers shown in Table 1). PCR amplification with DNA from *Musa acuminata* cv. Grande Naine (GN) was performed using primers BadnaFP and BadnaRP described above, and the product was cloned. Nine clones (FP2, FP4, FP6, FP14, FP19, FP20, FP22, FP26, and FP28) were sequenced in both orientations with primers M13F and M13R (GenBank accession numbers shown in Table 1). An additional BSGFV EPRV sequence was retrieved from the bacterial artificial chromosome (BAC) clone MBP_71C19 (GenBank AP009325) made from the genome of *Musa balbisiana* cv. PKW (Gayral et al. 2008).

A third category (labeled 'BSUgV') contained sequences of undetermined status (PRV or EPRV). This dataset of 105 sequences was obtained from Harper et al. (2005). These authors screened mainly *Musa acuminata* plants from BSV epidemics in Uganda and used a degenerate immuno-capture (IC)-PCR approach (see section Discussion). All available BSUgV sequences were downloaded from GenBank and were named according to information provided by GenBank.

In this study, total plant and viral genomic DNA was extracted from leaf tissue following a previously described method (Gawel and Jarret 1991). PCRs were carried out with 5–20 ng of DNA, 20 mM Tris–HCl (pH 8.4), 50 mM KCl, 0.1 mM of each dNTP, 1.5 mM $MgCl_2$, 10 pmol of each primer, and 1 U *Taq* DNA polymerase (Eurogentec, Seraing, Belgium) in 25 µl. PCRs were performed by first heating at 94°C for 4 min, followed by 35 cycles at 94°C for 30 s, 51–55°C for 30 s, and 72°C for 1 min/kbp, and one elongation cycle at 72°C for 10 min. PCR products were cloned into pGEM-T Easy (Promega, Madison, WI) or TOPO-TA (Invitrogen, Carlsbad, CA) vectors according to the manufacturer's instructions. Plasmid DNA was extracted with Wizard^R SV plus plasmid DNA purification system (Promega, Madison, WI) according to the manufacturer's instructions. Sequencing was performed by Cogenics Genome Express SA (Grenoble, France).

Phylogenetic Inference

Sequences were aligned using ClustalW (Thompson et al. 1994) implemented in Bioedit (Hall 1999) and corrected manually when necessary.

The software DAMBE version 4.5.20 (Xia and Xie 2001) was used to detect substitution saturation in each of the six alignments following a previously described method (Xia et al. 2003). For this purpose, the percentage of invariant sites was first estimated by DAMBE with a Poisson + Inv. distribution with the default parameters. For this analysis, the sequence of the outgroup *Taro bacilliform virus* (TaBV) was excluded from the alignments. The expected saturation index was given for asymmetric tree topology and estimated for 16 OTU after 500 replicates.

Before inferring phylogenies, we used Modeltest 3.7 (Posada and Crandall 1998) to choose the evolutionary model that best fitted our data using the Akaike informative criterion (AIC), since AIC has been shown to be better than hierarchical likelihood ratio tests to select a model (Posada and Buckley 2004). For each alignment, the best model and associated parameters were then used to infer tree topologies by maximum likelihood using PAUP 4.0b10

Table 1 Categories of sequences used in this study

Category	Sequence name	Clone name and GenBank accession number	Reference
BSUgV	BSUgBV	BSUgBV115 (AJ968463)	Harper et al. (2005)
	BSUgCV	BSUgCV114 (AJ968464)	
	BSUgDV	BSUgDV521 (AJ968465)	
	BSUgEV	BSUgEV112 and -523 (AJ968466, AJ968467)	
	BSUGFV	BSUGFV113 (AJ968469)	
	BSUgGV	BSUgGV532 (AJ968471)	
	BSUgHV	BSUgHV221 (AJ968472)	
	BSUgIV	BSUgIV14, -15, -16 (AJ968475 to AJ968477), BSUgIV172, -173, -191 (AJ968492 to AJ968494), BSUgIV193, -210, -212, -31, -36 (AJ968495, AJ968481, AJ968483 to AJ968485), BSUgIV421, -422, 423, -432, -436 (AJ968496 to AJ968500) BSUgIV45, -51, -53, -56, AJ968488, AJ968489 to AJ968491)	
	BSUgJV	BSUgJV293 and -296 (AJ968502, AJ968503)	
	BSUgKV	BSUgKV81, -82, -94 (AJ968504, AJ968505, AJ968507)	
	BSUgLV	BSUgLV222, -271, -283, -285, -32, -333, -344, -346 (AJ968517 to AJ968520, AJ968508, AJ968523 to AJ968525), BSUgLV610, -611, -73, -74, -76, -83, -84 (AJ968510 to AJ968516)	
	BSUgMV	BSUgMV102, -104, -132, -135, -143, -162 (AJ968526, AJ968527, AJ968529 to AJ968531, AJ968533), BSUgMV164, -165, -215, -263, -301, -302 (AJ968534, AJ968535, AJ968539 to AJ968542), BSUgMV321, -364, -365, -372, -373, -381, -382 (AJ968544, AJ968547, AJ968548, AJ968550, AJ968551), BSUgMV381, -382, -383, -511, -515, -516 (AJ968553 to AJ968558)	
	BSUgAV	BSUgAV445, 452,-456, -466, -472, -473, -481, -482 (AJ968454, AJ968455, AJ968457 to AJ968462)	
	BSUgGFV	BSUgGFV55, -542, -544, -545, -546, -548 (AJ968437 to AJ968442)	
	BSUgImV	BSUgImV11, -26, -232, -391, -492, -496 (AJ968444, AJ968446, AJ968448 to AJ968451)	
	BSUgOLV	BSUgOLV154, -171, -181, -182, -231, -244, -284, -311, -322 (AJ968422 to AJ968430), BSUgOLV342, -42, -43 (AJ968432 to AJ968434), BSUgOLV43 and -44 (AJ968419, AJ968420)	

Category	Banana host species	Clone name and GenBank accession number	Reference
EPRV	*Musa schizocarpa* (SS)	Shiz2, Shiz3, Shiz25, Shiz14, Shiz23, Shiz24 (Accession number AY189378 to AY189383)	Geering et al. (2005a)
	Musa acuminata subsp. *banksii* (Lescot et al.)	Bank1, Bank10, Bank11, Bank13, Bank14, Bank17, Bank19, Bank6, Bank8 (AY189384 to AY189392), Bank9 (AY452278)	
	Musa acuminata subsp. *burmannicoides* (Lescot et al.) cv. 'Calcutta 4'	Cal12, Cal13, Cal1, Cal22, Cal27, Cal30, Cal34, Cal6, Cal8, Cal22t (AY189444 to AY189453)	
	Musa acuminata subsp. *malaccensis* (Lescot et al.)	Mal10, Mal11, Mal15, Mal22, Mal26, Mal3, Mal6, Mal8 (AY189393 to AY189400)	
	Musa balbisiana cv. 'Pisang Batu' (BB)	Bat10, Bat19, Bat20, Bat21, Bat24, Bat25, Bat27, Bat2, Bat31, Bat34, Bat36, Bat4, Bat5, Bat6, Bat8, Bat9 (AY189420 to AY189435)	
	Musa balbisiana cv. PKW (BB)	PKW12, PKW16, PKW18, PKW23, PKW32, PKW36, PKW8, PKW9 (AY189436 to AY189443)	
	cv. 'Obino l'Ewai' (genotype AAB)	OBLE15, OBLE17, OBLE1, OBLE21, OBLE24, OBLE2, OBLE32, OBLE34, OBLE35, OBLE36, OBLE37, OBLE3, OBLE4, OBLE5, OBLE7, OBLE8, OBLE13t, OBLE1t, OBLE22t (AY189401 to AY189419)	
	cv. 'Klue Tiparot' (genotype ABB)	KT11, KT23 (AY452259 and AY452260), KT30, KT31, KT32, KT36, KT37, KT38, KT42, KT51, KT6, KT9 (AY452262 to AY452271)	
	Musa acuminata cv. 'grande Naine' (AAA)	FP2, FP4, FP6, FP14, FP19, FP20, FP22, FP26, FP28 (EU908850 to EU908858)	This study
	Musa balbisiana cv. PKW (BB)	PKW514, PKW515 (EU908849, EU919516)	This study
	Musa balbisiana cv. PKW (BB)	EPRVGF	Gayral et al. (2008)

Table 1 continued

Category	Species name	Accession number
PRV	*Taro bacilliform virus* (TabV)	AF357836
	Citrus yellow mosaic virus (CMBV)	AF347695
	Cacao swollen shoot virus (CSSV)	L14546
	Kalanchoe top-spotting virus (KTSV)	AY180137
	Commelina yellow mottle virus (ComYMV)	X52938
	Sugarcane bacilliform Mor virus (SCBMV)	M89923
	Banana streak Obino l'Ewai virus (BSOLV)	NC_003381
	Banana streak Mysore virus (BSMysV)	NC_006955
	Banana streak Acuminata Vietnam virus (BSAcVnV)	AY750155
	Banana streak Acuminata Yunnan virus (BSAcYuV)	DQ092436
	Banana streak Imove virus (BSImV)	Unpublished data
	Banana streak Goldfinger virus (BSGFV)	AY493509
	Banana streak Cavendish virus (BSCavV)	This study (EU908859)

(Swofford 2002). Branching supports were assessed by performing 500 bootstrap replicates using PHYML (Guindon and Gascuel 2003).

At the same time, a Bayesian approach was also used to confirm the topologies of each alignment previously inferred with the ML approach, using the software MrBayes 3.1.2 (Huelsenbeck and Ronquist 2001). Each run was performed with five chains and 10^6 generations using the default priors of the GTR model. Bayesian posterior probabilities were calculated from majority-rule consensus of trees sampled every 20 generations once the Markov chains had become stationary (determined by empirical checking of likelihood values).

Estimation of Evolution Rate and Selective Pressures

The phylogenetic trees were labeled to differentiate the branches leading to 'PRV,' to 'EPRV' or to 'BSUgV' sequences. Since episomal viral sequences can integrate, and, conversely, EPRV can be activated, the status of a taxa or lineage can change over time, i.e., along the branches of the phylogenetic tree. The real status of a sequence is therefore known only at the time of its sampling. To minimize the biases caused by the wrong classification of sequences, only the terminal branches were labeled. Furthermore, our sampling showed sufficient diversity to provide many short terminal branches.

Variations in synonymous and nonsynonymous substitution rates for each sequence category were analyzed using the *branch models* of the codeml program implemented in PAML 3.15 (Yang 1997, 1998). To compare the d_N/d_S ratio between 'PRV' and 'EPRV' for instance, two models were built. A null model (M_G) assuming a common d_N/d_S ratio for 'PRV' and 'EPRV' terminal branches, plus a second ratio for the remaining branches, i.e., the internal branches and the 'BSUgV' branches' (Table 2). M_G was compared to an alternative model (M_H) assuming one ratio for 'PRV' and one for 'EPRV,' and one for the remaining branches. The likelihood ratio of the two models to be compared (M_G vs. M_H) tested whether the alternative model fitted the data significantly better than the null hypothesis: twice the difference in log likelihood between the two models was compared with a χ^2 distribution with n degrees of freedom, n being the difference between the numbers of parameters of the two models. Details of each nested models used in this study and the corresponding question addressed are shown in Table 2. Stop codons were found in several EPRV sequences and were removed prior to PAML analysis. Inserted nucleotides were removed; deleted nucleotides and the third base of substituted stop codons were coded as unknown. Estimations of variations of d_N/d_S ratio between endogenous and episomal viral sequences were used as a proxy for investigating the selective forces occurring before and after BSV integrations.

To check that an increase in the nonsynonymous/synonymous (d_N/d_S) ratio was only due to a relaxation of selective constraints, rather than to the effect of positive selection, the latter was tested by the *site models* of codeml implemented in PAML (Yang 1998; Yang and Nielsen 1998). The neutral model (M_7) uses a discrete beta distribution [range (0, 1)] to model different d_N/d_S ratios among sites. The alternative model M_8 assumes a supplementary class of codons with $d_N/d_S > 1$. M_7 and M_8 were compared with likelihood ratio tests (LRT).

The program baseml of PAML implements local clock models which assume that branches of the tree can be divided into several rate groups (Yang and Yoder 2003; Yoder and Yang 2000). As for the comparison of the d_N/d_S ratio, the last branches of the tree were labeled in categories to test if PRV and EPRV had the same or a distinct

Table 2 Models used in this study and biological questions addressed by comparing models

Model	Estimations/detections	Schematic definition of the model	Model comparison	Question addressed
M_A	d_N/d_S ratio	d_N/d_S internal branches, d_N/d_S PRV = d_N/d_S BSUgV	M_A vs. M_B	Are the BSUgV and PRV selective regimes distinct?
M_B	d_N/d_S ratio	d_N/d_S internal branches, d_N/d_S PRV, d_N/d_S BSUgV		
M_C	d_N/d_S ratio	d_N/d_S internal branches, d_N/d_S EPRV = d_N/d_S BSUgV	M_C vs. M_D	Are the BSUgV and EPRV selective regimes distinct?
M_D	d_N/d_S ratio	d_N/d_S internal branches, d_N/d_S EPRV, d_N/d_S BSUgV		
M_E	Relative evolution rate (R)	R internal branches, R PRV = R EPRV	M_E vs. M_F	Are the BSUgV and PRV evolution rates distinct?
M_F	Relative evolution rate (R)	R internal branches, R PRV, R EPRV		
M_G	d_N/d_S ratio	d_N/d_S internal branches, d_N/d_S PRV = d_N/d_S EPRV	M_G vs. M_H	Are the EPRV and PRV selective regimes distinct?
M_H	d_N/d_S ratio	d_N/d_S internal branches, d_N/d_S PRV, d_N/d_S EPRV		
M_7	Positive selection	PAML site-model with d_N/d_S [0–1]	M_7. vs M_8	Do BSUgV evolve under positive selection?
M_8	Positive selection	PAML site-model with d_N/d_S [0–1] + $d_N/d_S > 1$		

substitution rate relative to the rate observed in the rest of the tree. Neutral models (M_E) assuming a first common rate for 'PRV' and 'EPRV' branches, plus a rate for the other branches (M_E: two-ratio models) were compared by LRTs with alternative models (M_F) assuming a rate for the branches of the category 'EPRV', a rate for 'PRV,' and a rate of the other branches (M_F: three-ratio models, see Table 2). A change in molecular evolution rate was examined between BSV and BSV EPRVs to propose an evolutionary scheme of integration events.

Results

Analysis of the Phylogenetic Signal

To check if our alignments were suitable for phylogenetic studies, we tested for the presence of substitution saturation with the program DAMBE. It is assumed that phylogenetic information is mainly lost when the observed saturation index is equal to or more than half full substitution saturation (Xia 1999). Expected saturation indices assuming half the full substitution saturation were estimated and compared to the observed saturation indices. Saturation is detected when the observed indices are higher than the expected indices. None of the six alignments used in this study (see below, and alignments in supplementary data) showed signs of saturation ($p < 0.03$, Table 3), thereby validating our dataset for phylogenetic analyses.

Overall Phylogeny of BSVs

To get an overview of the badnavirus phylogeny, a first phylogeny was inferred from 25 sequences chosen as a representative sample of the diversity of BSV sequences and close badnavirus species (named *overall PRV* alignment) and is shown in Fig. 1. Three deeply rooted groups were distinguished and named groups 1, 2, and 3. These groups were supported by higher bootstrap values in phylogenies with additional sequences (see Figs. 2, 3, 4). However, the order of emergence of the three groups is still not clear as suggested by the basal trifurcation.

Table 3 Detection of saturation substitution with the program DAMBE

Alignment[a]	Observed saturation index (Weiss)	Expected saturation index (Iss.cAsym)	T	DF	*p*-value[b]
Overall PRV	0.458	0.518	2.16	473	0.0317
Group 1	0.445	0.518	2.58	434	0.0102
Group 2	0.355	0.516	6.24	506	0.0000
Group 3	0.373	0.517	5.54	427	0.0000
Group 1 + 2	0.448	0.519	2.15	500	0.0319
Main BSUgV	0.435	0.517	3.11	427	0.0020

[a] Alignments detailed in the Results section

[b] The statistical significance of the difference between observed and expected saturation indexes was assessed with a two-tailed test

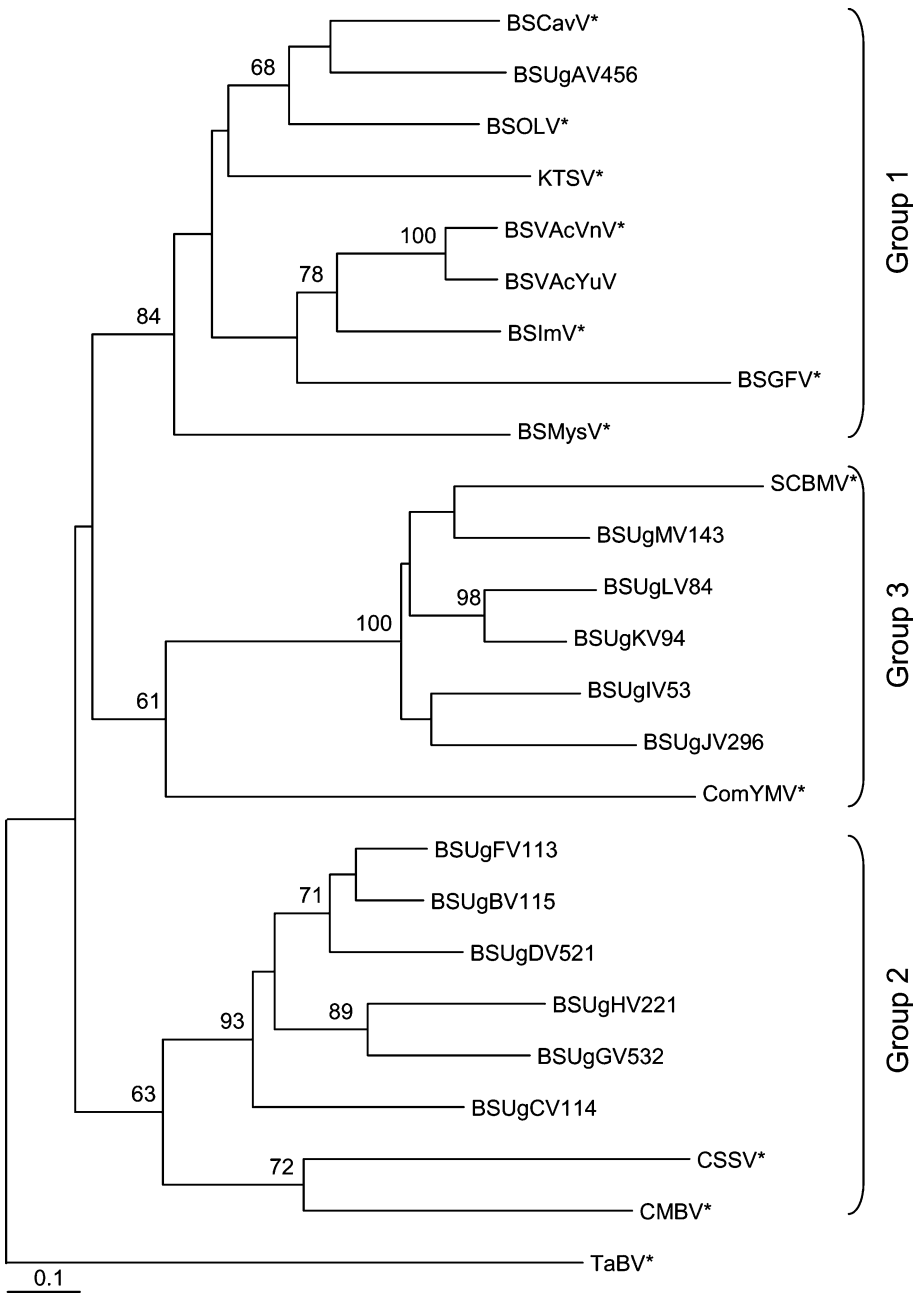

Fig. 1 Maximum likelihood phylogeny of overall badnaviruses in RT/RNase H region. Bootstrap values of 500 replicates are given above nodes when >60%. Fully described and known episomal viruses are indicated with an *asterisk*

All previously described species of BSV clustered in group 1: BSV species Obino l'Ewai (BSOLV) (Harper and Hull 1998), BSV species Golfinger (BSGFV) (Gayral et al. 2008), BSV species Mysore (BSMysV) (Geering et al. 2005b), BSV species Imové (BSImV) (Gayral et al., unpublished data), and BSV species acuminata Vietnam (BSAcVNV) (Lheureux et al. 2007). Three subgroups were observed in group 1. The first subgroup contained BSMysV species; the second subgroup contained BSOLV and BSV species Cavendish (BSCavV), the newly discovered sequence BSUgAV, and the kalanchoe top-spotting badnavirus (KTSV) isolated on *Kalanchoe blossfeldiana*; and the third subgroup contained banana-infecting species only: BSGFV, BSV species Yunnan (BSAcYuV), BSAcVNV, and BSImV.

Groups 2 and 3 contained closely related badnavirus species, but mainly BSUgV sequences collected from infected banana cultivars during Uganda epidemics (Harper et al. 2005). Group 2 contained seven distinct taxa of BSUgV as well as *Citrus mosaic bacilliform virus* (CMBV) and *Cacao swollen shoot virus* (CSSV), which infect citrus and cocoa (*Theobroma cacao*), respectively, in a basal position. Group 3 contained both the other BSUgV taxa (from I to M) and *Sugarcane bacilliform mor virus*

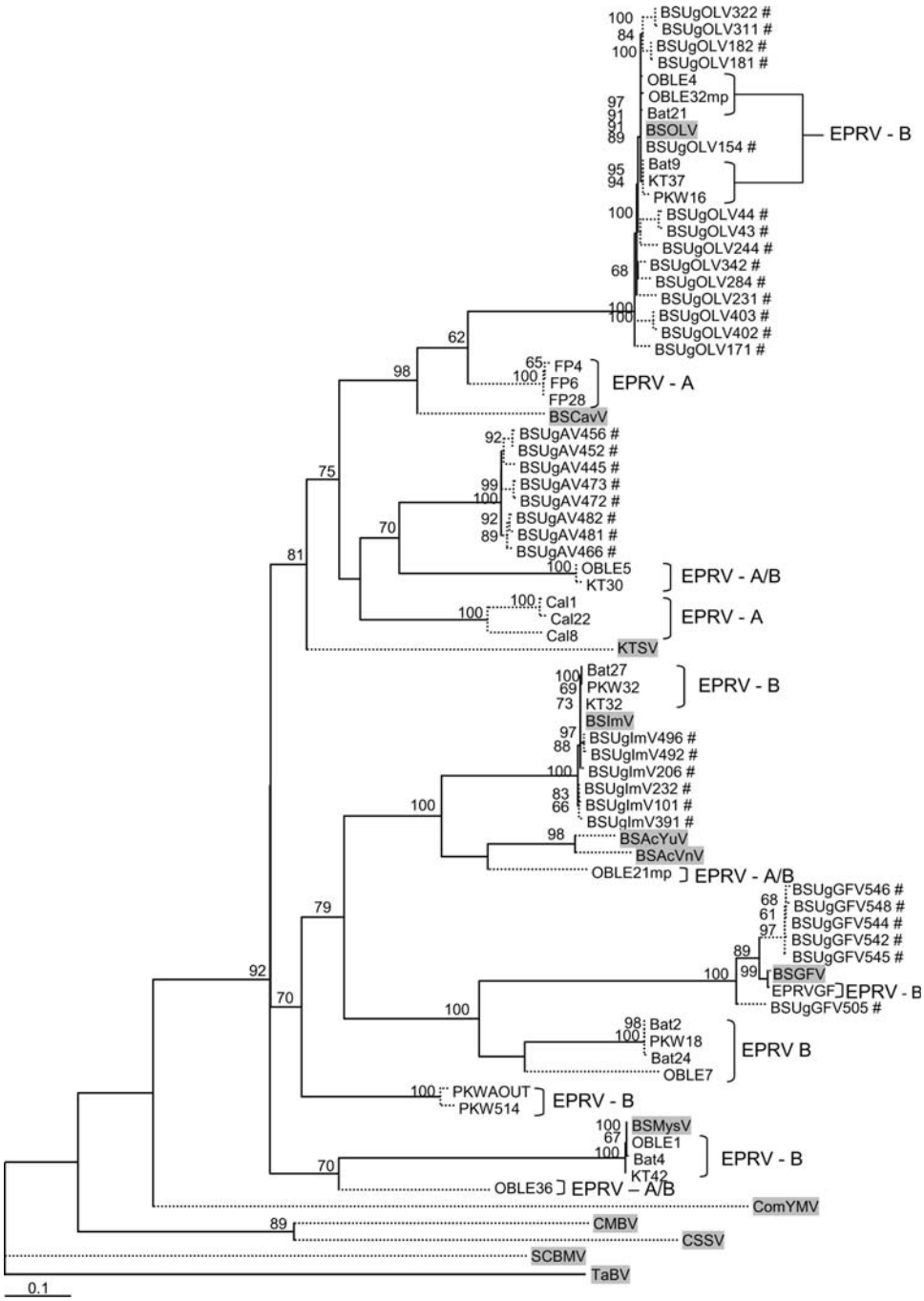

Fig. 2 Maximum likelihood phylogeny of episomal and endogenous BSV of group 1. Phylogeny is based on a 540 bp alignment of RT/RNase H viral region. Bootstrap values of 500 replicates are given when >60%. Other sequences of groups 2 and 3 are given as outgroups. BSUgV sequences are indicated by a *hash sign* (#). Episomal sequences of both BSV and badnaviruses are *shaded*. EPRV-A and EPRV-B are sequences integrated in the *Musa acuminata* genome (denoted *A*) and the *M. balbisiana* genome (denoted *B*), respectively. EPRV-A/B sequences are found in interspecific A × B *Musa* genotypes and cannot be assigned to a particular genome when not close to a given phylogenetic group

(SCBMV), which infect sugarcane (*Saccharum officinarum*). *Commelina yellow mottle virus* (ComYMV), which infects *Commelina* sp., was in basal position in this group.

Origin and Evolution of Endogenous BSV

A ML phylogeny based on the initial dataset containing all 217 sequences was first reconstructed and the same tree topology as in Fig. 1 was obtained (data not shown). To facilitate further analyses, this initial alignment was split into three parts according to the three major phylogenetic groups. *Alignment 1*, *alignment 2*, and *alignment 3* contained 76, 86, and 72 sequences, respectively, and encompassed all the sequences that belong to groups 1, 2, and 3, respectively, as well as sequences from the other two groups as outgroups. All sequences were labeled as either 'PRV' for episomal viruses, 'EPRV' for endogenous sequences or 'BSUgV' for sequences from Uganda with unclear endogenous or episomal status. We studied the distribution of integration events in the *Musa* genome by

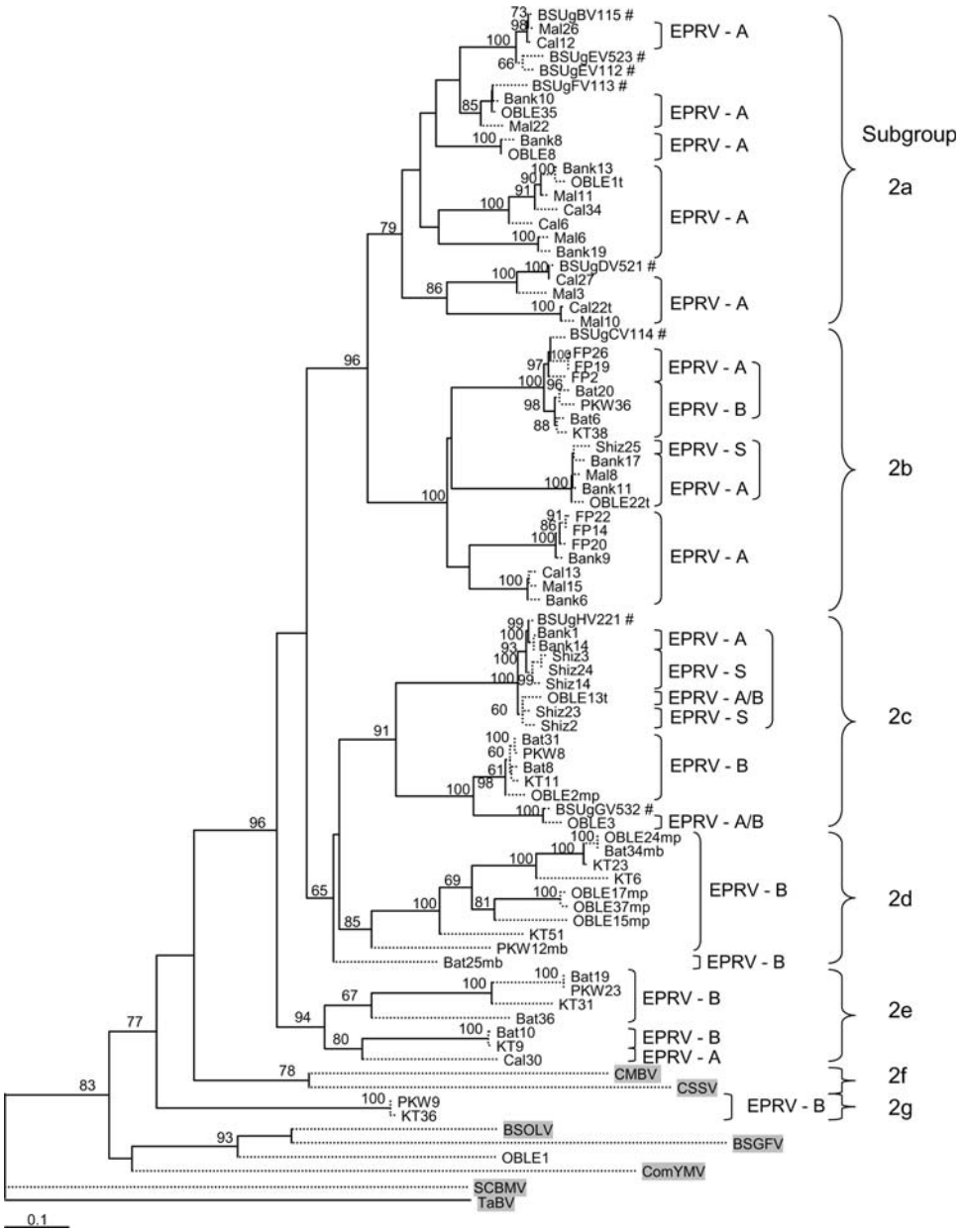

Fig. 3 Maximum likelihood phylogeny of episomal and endogenous PRV of group 2. The phylogeny was established with a 540 bp alignment of the RT/RNase H viral region. Bootstrap values of 500 replicates are given when >60%. Other sequences of groups 1 and 3 are given as outgroups. BSUgV taxa are indicated by a *hash sign* (#). Episomal BSV and episomal badnaviruses are *shaded*. EPRV-A, EPRV-B, and EPRV-S are sequences integrated in the *Musa acuminata* (A), *M. balbisiana* (B), and *M. schizocarpa* (S) genomes, respectively. EPRV-A/B sequences are found in interspecific A × B *Musa* genotypes and cannot be assigned to a particular genome when not close to a given phylogenetic group

distinguishing between two independents events. We defined an independent event as one or more EPRV sequences clustering in a single and well-supported phylogenetic group and separated from other EPRVs by episomal sequences.

Group 1 (Fig. 2) contained EPRV, PRV as well as BSUgV sequences. Except for BSUgV species A (denoted BSUgAV) that formed a new phylogenetic group that was distinct from any other known BSV species, all BSUgV sequences corresponded to previously described BSV species such as BSOLV, BSImV, and BSGFV. We recorded at least 10 independent integration events in this group, the majority (6/10) were restricted to the B genome and are denoted as EPRV-B in Fig. 2. They occurred in diploid *M. balbisiana* genotypes (BB) such as cv. PKW and cv. Pisang batu, but also in interspecific genotypes such as cv. Klue tiparot (ABB) and cv. Obino l'Ewai (AAB). At least two integrations were found in A × B interspecific hybrids. Since they did not cluster near EPRV-A or EPRV-B integration, it was not possible to attribute them to the A genome or B genome. Finally, two integrations were specific to the A genome and appeared in *Musa acuminata* (Lescot et al. 2008) only. It is not clear whether the following sequences integrated in the A genome (clones Cal1, Cal22, and Cal8) and in the genome A or B (clones OBLE5 and KT30) derived from the same integration event or from two independent events, since they are not separated by episomal sequences. Nevertheless, the relatively deep

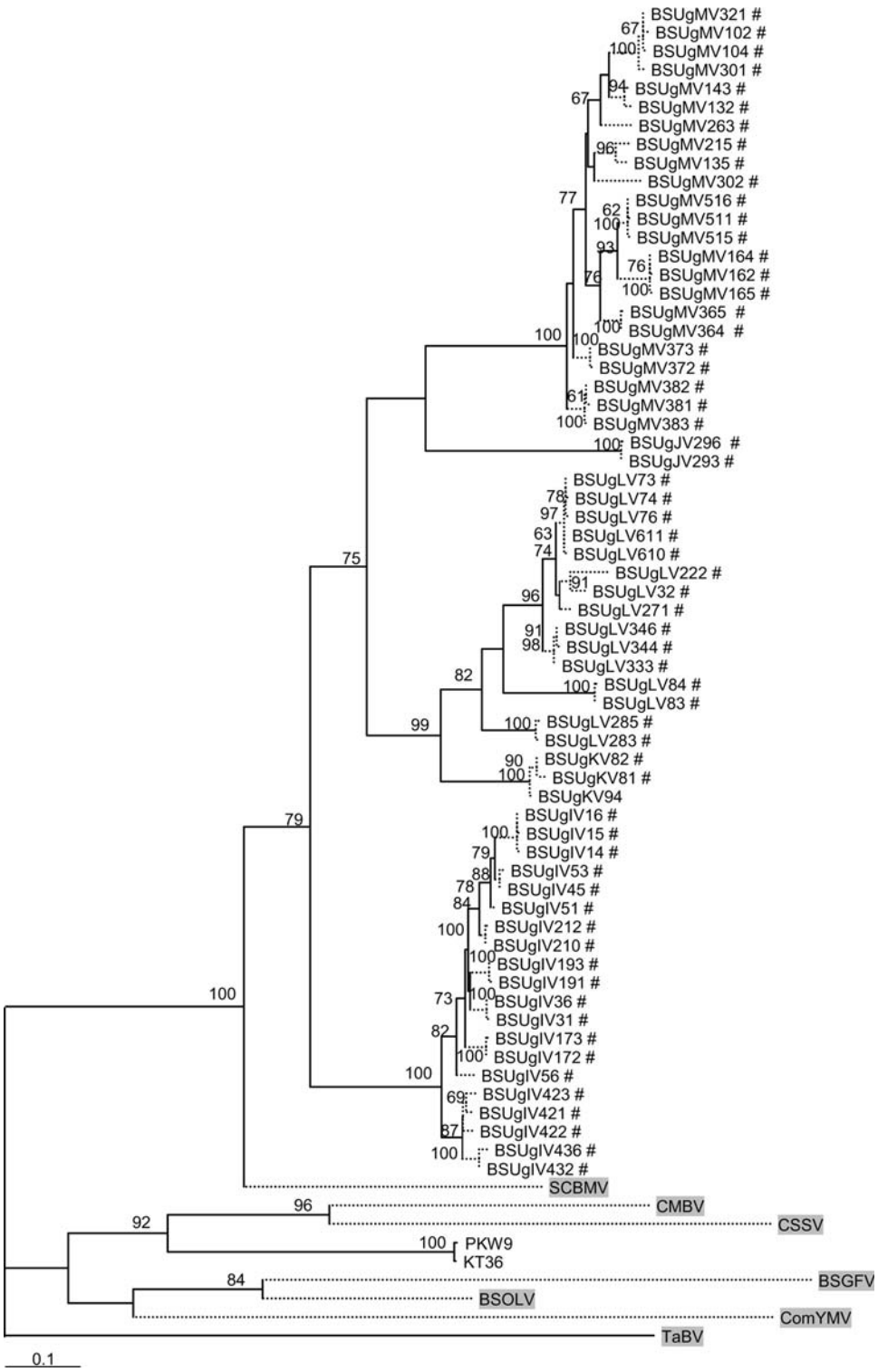

Fig. 4 Maximum likelihood phylogeny of episomal and endogenous PRV of group 3. The phylogeny was established with a 540 bp alignment of the RT/RNase H viral region. Bootstrap values of 500 replicates are given when >60%. Other sequences of groups 1 and 2 are given as outgroups. BSUgV taxa are indicated by a *hash sign* (#). Episomal sequences of both BSV and badnaviruses are *shaded*

divergence observed between the two groups (Cal8-KT30: 0.61 substitution/site) suggests that two independent integrations occurred. Figure 2 shows that EPRVs related to BSOLV, BSImV, BSGFV, and BSMysV are associated with the B genome, whereas EPRVs related to BSCavV, BSAcYuV, and BSAcVNV are more associated with the A genome. EPRVs corresponding to unknown viruses were observed in both A and B genomes. Furthermore, we did not observe any EPRV belonging to Group 1 in the *Musa schizocarpa* genome.

Group 2 contained the largest proportion of EPRVs analyzed in this study as well as seven of the newly

discovered BSUgV species (namely species B to H) that infect bananas. Seven subgroups were defined and named 2a to 2g (Fig. 3). Two subgroups diverged early during the evolution of group 2. The first, subgroup 2g, contained two EPRVs (PKW9 and KT36) found in the B genome. The second corresponded to two badnaviruses that do not infect *Musa*: CMBV and CSSV (subgroup 2f). It is not clear whether subgroup 2g diverged first, since even if the node was well supported by bootstrap value in the 'Alignment 2' phylogeny (77%), it was much less supported (bootstrap value = 34%) in the Bayesian approach (branch trifurcation; data not shown) as well as in the ML phylogeny in the larger dataset 'Alignment 1 + 2' (data not shown). The latter alignment merged *alignment 1* and *alignment 2* in a single dataset of 154 sequences and encompassed the only two groups containing EPRV sequences; group 3 was indeed free of EPRV (see below).

The count uncertainty of integration events for group 2 was expected to be higher than for group 1 since fewer reliable episomal virus clustered in this group. However, several BSUgV sequences that separate EPRV sequences were probably episomal viruses. Furthermore, the major divergence between groups of EPRV sequences often enables discrimination between integration events (see below). Sequences of group 2 corresponded to endogenous sequences within the three banana species studied: *M. balbisiana*, *M. acuminata*, and *M. schizocarpa* (genome S). In several subgroups, EPRVs sequences belonged to a single *Musa* species: subgroup 2a integrated the A genome only, subgroups 2d and 2g integrated the B genome only, suggesting that they originated from independent integration events. All the sequences derived from a single integration event are therefore paralogous. The situation was not the same for subgroups 2b, 2c, and 2e, since sister groups were described in different *Musa* species. In subgroup 2e, one group integrated the B genome and its sister group the A genome, but the two sister groups showed relatively high divergence (0.4 substitution/site). However, in subgroups 2b and 2c, the topology was similar: the two sister groups integrated in genomes A and B (subgroup 2b) and in genomes A and S (subgroups 2b and 2c), but in this case the divergence between the sister groups was much lower (0.08, 0.04, and 0.03 substitution/site, respectively). In three cases, the integration events most likely occurred before the speciation of the *Musa* genus. Sequences of sister groups are thus orthologous. This result indicates the occurrence of ancient integrations events. Altogether, at least 17 independent integration events were found in group 2: seven were specific to genome A, six to genome B, one was undetermined (either in A or in B), one was common to A and B, and two were common to the *Musa* genomes A and S.

Figure 4 shows the ML phylogeny of group 3. Surprisingly, this group was free of any described EPRV and was mainly composed of BSUgV sequences. A subgroup diverged first and is represented by the only SCBMV sequence, a virus that infects sugarcane. The three other subgroups were composed of BSUgV sequences (species I to M). For all alignments, Bayesian reconstructions produced congruent topology with the ML inferences at the level of general topology as well as for terminal branching of trees (data not shown).

Rate of Evolution and Selection of BSV and EPRV

Status of BSUgV Sequences from Uganda Epidemics

The phylogeny of endogenous and episomal sequences revealed that EPRV sequences did not form a limited phylogenetic group, but were dispersed in two clades (groups 1 and 2). It was consequently impossible to assign the category of a given sequence solely by its phylogenetic position. We addressed this question for BSUgV sequences distributed in the three groups and whose category was not clearly known. Their molecular evolution pattern was studied to determine if they evolved as PRV, or as EPRV sequences. Table 4 shows for each phylogenetic group the estimations and comparison by LRT of the d_N/d_S ratio and evolution rate of the terminal branches of different sequences categories: 'PRV,' 'EPRV,' and 'BSUgV.' For alignments corresponding to groups 1, 2, 3 and the combination of 1 and 2, BSUgV terminal branches had a d_N/d_S ratio that was significantly different from that of the PRV terminal branches. The d_N/d_S ratio was 30 times higher for BSUgV than PRV in group 2, and 4 times higher in groups 1 and 3. To test if this increase in the d_N/d_S ratio could be the result of positive selection acting on some codons rather than a relaxation of selective constraints, we searched for signals of positive selection in the BSUgV sequences using PAML. We defined a new alignment containing 41 of the 105 published BSUgV sequences representative of the genetic diversity of BSUgV observed in the three phylogenetic groups, as alignment *Main BSUgV*. No positive selection was found with this new dataset (Table 4: M_7 vs. M_8, $p = 1$). Furthermore, the d_N/d_S ratio of BSUgV branches did not significantly differ from those of EPRV branches in all phylogenetic groups. BSUgV sequences contained a molecular evolution signal closer to EPRV than to PRV, and this result was mostly observed in group 2. It is therefore possible that endogenous sequences exist among BSUgV sequences previously defined as PRV sequences. BSUgV sequences were therefore retained in our datasets since they obviously contributed to BSV phylogeny, but were categorized as neither EPRV nor PRV in subsequent steps of the molecular evolution analysis.

Table 4 d_N/d_S ratio and likelihood ratio tests of BSUgV terminal branches

Comparison	Dataset	Model	LnL	np	d_N/d_S internal branches	d_N/d_S PRV	d_N/d_S EPRV	d_N/d_S BSUgV	2dLnL	df	*p*
BSUgV vs. PRV	*Alignment 1*	M_A	−9620.48	152	0.040	0.042	–	$=d_N/d_S$ PRV	23.63	1	1.13E−06***
		M_B	−9608.66	153	0.040	0.018	–	0.079			
	Alignment 2	M_A	−13608.87	172	0.061	0.052	–	$=d_N/d_S$ PRV	62.42	1	2.78E−15***
		M_B	−13577.66	173	0.061	0.009	–	0.283			
	Alignment 3	M_A	−8029.23	144	0.035	0.052	–	$=d_N/d_S$ PRV	10.19	1	1.41E−03**
		M_B	−8024.13	145	0.034	0.014	–	0.063			
	Alignment 1 + 2	M_A	−19285.55	308	0.056	0.054	–	$=d_N/d_S$ PRV	60.25	1	8.330E−15***
		M_B	−19255.42	309	0.055	0.016	–	0.127			
BSUgV vs. EPRV	*Alignment 1*	M_C	−9604.32	152	0.027	–	0.081	$=d_N/d_S$ EPRV	0.00	1	$9.61E-01^{NS}$
		M_D	−9604.32	153	0.027	–	0.081	0.082			
	Alignment 2	M_C	−13518.25	172	0.029	–	0.188	$=d_N/d_S$ EPRV	3.07	1	$7.99E-02^{NS}$
		M_D	−13516.71	173	0.029	–	0.175	0.291			
	Alignment 1 + 2	M_C	−19189.49	308	0.028	–	0.142	$=d_N/d_S$ EPRV	0.46	1	$4.98E-01^{NS}$
		M_D	−19189.26	309	0.028	–	0.148	0.130			
Positive selection	*Alignment Main BSUgV*	M_7	−6576.11	82					0.00	2	1^{NS}
		M_8	−6576.11	84							

LnL Log-likelihood of the model, *np* number of parameters of the model, d_N/d_S *other* parameter for all branches except those of PRV and EPRV terminal branches, *2dLnL* twice the likelihood of the two compared models, *df* number of degree of freedom, *NS* not significant

Significant at the ** 1% level and *** 0.1% level

Evolution of EPRV and PRV Terminal Branches

The evolutionary pattern between terminal branches of EPRV and those of PRV was compared using LRT. Table 5 shows the estimates of d_N/d_S ratio and relative rates of evolution of terminal branches of PRVs and EPRVs, for each phylogenetic group in which integration occurred (groups 1 and 2). In all phylogenetic groups, the speed of evolution of EPRV terminal branches was significantly lower than that of PRV branches, with a *r*EPRV/*r*PRV factor of about 0.6 for group 1 and group 2.

Table 5 d_N/d_S ratio, evolution rates estimates, and likelihood ratio tests of PRV and EPRV terminal branches

Parameter	Dataset	Model	LnL	np	Branch label			2dLnL	df	*p*
					Internal	PRV	EPRV			
R	*Alignment 1*	M_E	−10176.41	77	1	0.942	=rPRV	7.91	1	4.91E−03***
		M_F	−10172.45	78	1	1.090	0.702			
	Alignment 2	M_E	−14202.95	87	1	0.975	=rPRV	20.00	1	7.75E−06***
		M_F	−14192.95	88	1	1.338	0.739			
	Alignment 1 + 2	M_E	−20059.24	155	1	0.940	=rPRV	13.23	1	2.76E−04***
		M_F	−20052.63	156	1	1.168	0.795			
d_N/d_S	*Alignment 1*	M_G	−9620.40	152	0.042	0.039	$=d_N/d_S$ PRV	19.04	1	1.28E−05***
		M_H	−9610.88	153	0.040	0.018	0.071			
	Alignment 2	M_G	−13575.87	172	0.038	0.115	$=d_N/d_S$ PRV	74.43	1	0***
		M_H	−13538.66	173	0.037	0.003	0.165			
	Alignment 1 + 2	M_G	−19260.56	308	0.041	0.089	$=d_N/d_S$ PRV	91.66	1	0***
		M_H	−19214.73	309	0.039	0.015	0.133			

LnL log-likelihood of the model, *np* number of parameters of the model, *Branch label* = *other* all branches except from PRV and EPRV terminal branches, *2dLnL* twice the likelihood of the two models compared, *df* number of degree of freedom

*** Significant at the 0.1% level

Furthermore, terminal branches of EPRVs evolved under much lower selective pressure than terminal branches of PRVs (Table 5). The d_N/d_S ratio of EPRVs was 4 and 55 times higher than those of PRV in groups 1 and 2, respectively (e.g., Group 2: d_N/d_S PRV = 0.003, d_N/d_S EPRV = 0.165). This strong relaxation of selective constraints was similar to that observed between BSUgV and PRV sequences (Table 4), confirming the probable endogenous nature of several BSUgV sequences. Additionally, *in silico* translation of the 99 EPRV sequences used in this study confirmed that 43 of them showed signs of pseudogenization, such as substitutions leading to an in-frame stop codon or indels leading to a premature stop codon.

Comparison Between Host Plant and Virus Substitution Rates

We used existing data on *Hepatitis B virus* (HBV, *Hepadnaviridae* or animal pararetroviruses) and Retroviruses, as a proxy for PRV substitution rate. First, *Hepadnaviridae* are closely related to the *Metaviridae* group that includes *Caulimoviridae* (Malik and Eickbush 2001). Second, HBV, retroviruses and PRVs use the same polymerase (reverse transcriptase—RT). It has been shown that the viral replication enzyme is an important determinant of evolutionary changes in viruses (Duffy et al. 2008), mostly because RTs do not have proofreading capabilities and are therefore more error prone than DNA polymerases (Flint et al. 2003). The substitution rate of HBV was estimated at 10^{-4}–10^{-5} substitutions per site per year (subs/site/year) (Zhou and Holmes 2007), and those of retroviruses ranges 10^{-3}–10^{-6} subs/site/year (Hanada et al. 2004; Jenkins et al. 2002). We therefore hypothesized that the substitution rate of plant pararetroviruses, and thus episomal BSV, range between 10^{-3} and 10^{-6} subs/site/year.

In banana, the neutral evolution rate (synonymous substitution rate) of the nuclear genome was estimated at 4.5×10^{-9} subs/site/year (Lescot et al. 2008), and EPRV sequences should evolve at a comparable rate. We thus expected a difference of at least three orders of magnitude between the evolution rate of PRV sequences and those of EPRV sequences, but the difference recorded was at most a factor 1.8 (Table 5: R_{PRV}/R_{EPRV} = 1.338/0.739).

Discussion

Today's existing BSV phylogenies are focused either on episomal sequences or on endogenous BSV only. Our work combined for the first time all available episomal and endogenous badnavirus sequences that contribute to the genetic diversity of this genus. Since no genomic recombination between BSV species could be detected (E. Muller, unpublished data), and no substitution saturation was present in our dataset, we assume that the RT/RNaseH portion of ORF3 analyzed in this study provided a nonbiased picture of BSV phylogenetic relationships.

The BSV phylogeny obtained herein confirmed the existence of three distinct groups of BSVs, as previously suggested by Harper et al. (2005). Furthermore, we showed that BSVs are polyphyletic since several BSUgV sequences in groups 2 and 3 were closely related to viruses that do not naturally infect banana such as SCBMV, ComYMV, CSSV, and CMBV, and more distantly related to the known BSV group 1 that does infect bananas. The badnavirus phylogeny depicted in our study could thus be used to revise the taxonomy of this genus. The naming of endogenous BSV sequences refers for instance to *Banana streak virus* species, which may not be an appropriate name given the large molecular diversity of badnaviruses. The name Banana endogenous virus-X (BEV-X), proposed by Geering et al. (2005a), should be used as a tentative sequence name when the episomal virus is unknown or when the phylogenetic relationships are not clearly established.

Banana is a monocotyledon and the only natural host of BSV species belonging to the three phylogenetic groups (Jones 2000). KTSV, SCBMV, ComYMV, CSSV, and CMBV are viruses that all clustered close to BSV, but under natural conditions, these viral species are unable to infect banana (Fauquet et al. 2005). SCBMV infects sugarcane (monocotyledon), and CSSV is found in cocoa (dicotyledon). The phylogeny presented here suggests that banana was the host plant for the ancestor of BSVs and of the five above-mentioned badnaviruses, and that independent host shifts subsequently occurred during the evolution of these viral species. Viral host shifts are facilitated when two plant species are colonized by the same insect species that transmits the virus (Harper et al. 2005; Jones 2000). In many tropical countries, banana is often grown near sugarcane and the mealybug *Planoccocus citri*—a BSV vector—feeds on both plants (Jones 2000). Likewise, *P. citri* also feeds on citrus (Ben-Dov 1994), suggesting that this mealybug species might have contributed to the host shift of SCBMV and CMBV. Interestingly, different isolates of SCBV from sugarcane are still able to infect banana when agro-inoculated (Bouhida et al. 1993), suggesting either that the host shift was recent or that SCBV is a relatively generalist pathogen.

Harper et al. (2005) published sequences from Ugandan epidemics assumed to correspond to episomal viruses. The authors used IC-PCR and direct binding degenerate PCR to detect episomal viruses only. However, these methods often lead not only to co-amplification of the virus but also to co-amplification of EPRVs from residual *Musa* genomic DNA. In our study, we performed a systematic monitoring

of EPRV contamination through Multiplex-Immuno-Capture-PCR (Le Provost et al. 2006) to ensure that only episomal viruses are amplified. Our analysis of the molecular evolution of the BSUgV dataset clearly indicated the presence of contaminating EPRVs among them. This result underlines the need to check a priori the status of each sequence before its release in public sequence databases and its use in phylogenetic analyses. The other sequences used in this study were checked a priori and derived from both full-length circular genomes for the PRV dataset and from PCRs performed on genomic DNA of virus-free plant material for the EPRV dataset. This precaution remains valid in virology whenever endogenous counterparts exist in the host genome, such as retroviruses in animals (Bromham 2002; Weiss 2006), temperate phages in bacteria (Daubin and Ochman 2004), or geminiviruses in plants (Murad et al. 2004; Pal et al. 2007).

Combining endogenous and episomal sequences in the same phylogeny was useful to answer several evolutionary questions regarding the integration phenomenon in the banana genome. We first provided evidence of large-scale integrations of badnaviruses in the genome of at least three *Musa* species (*M. acuminata*, *M. balbisiana*, and *M. schizocarpa*). These integrations were frequent since 27 independent events were observed, and also recent since most of them occurred after the *M. acuminata*/*M. balbisiana* speciation, i.e., ca. 4.6 My ago, as estimated from the molecular evolution of zingiberales calibrated with paleontological data (Lescot et al. 2008). Only three integrations occurred before this date, suggesting a long co-evolution of the virus and its host plant. Our data did not support co-diversification of badnaviruses and banana, since sequences belonging to the three viral genetic groups were integrated in the genome of the three banana species studied. However, integration in *M. schizocarpa* appeared to be rare and concerned only group 2, but this might be due to insufficient ERPV sampling for this species. Furthermore, our analysis indicated that group 2 is mostly represented by endogenous sequences. Only eight BSUgV sequences belong to this group and were probably EPRVs too. This result either means that episomal viruses in group 2 exist at the present time but have not yet been described or that a lineage of episomal viruses existed that subsequently became extinct. The search for episomal group 2 viruses in several *Musa* genotypes and species is required to address this issue. Finally, the question regarding the lack of EPRVs in group 3 remains unanswered. A search for EPRVs homologous to the viral sequences that comprise group 3 in banana and other plants species considered as putative ancient host is required to trace the co-evolutionary history of this group.

Our phylogenetic analysis showed that the large diversity of endogenous badnaviruses in the *Musa* genome is the consequence of multiple independent integration events of numerous viral species. This result confirms previous experimental work based on single known BSV species in a limited number of banana genotypes. Fluorescent in situ hybridization (FISH) experiments and fingerprint analyses of *Musa* bacterial artificial chromosome (BAC) libraries showed that the *Musa* B genome harbors only a limited number of EPRV copies for a given viral species (Gayral et al. 2008; Harper et al. 1999; Iskra-Caruana, unpublished data). The genome of several *Solanaceae* species (*Petunia* sp., *Nicotiana* sp., *Solanum* sp.) harbors a different EPRV integration pattern from that of the banana genome. In these species, few EPRVs underwent large-scale amplification and reached hundreds to thousands copies in the host genome (Gregor et al. 2004; Mette et al. 2002; Richert-Poggeler et al. 2003; Staginnus et al. 2007). One possible explanation for this difference from banana is that BSV integrated the *Musa* genome much later, and further amplification of endogenous badnaviruses may not have occurred so far. Furthermore, recent integration events would also explain why the viral ORFs of at least four EPRVs in *Musa* are conserved (BSOLV in cv. Obino L'Ewai and cv. PKW, BSImV and BSGFV in cv. PKW), and therefore remain likely to induce infection after activation.

The comparison of evolutionary parameters in PRV and EPRV sequences yielded unexpected results. PRV sequences evolved more slowly than expected, and/or EPRV sequences evolved more rapidly than expected. Two nonexclusive phenomena could have modified the evolutionary parameters in such a way. The first possible explanation is that some branches labeled as EPRV correspond to primarily episomal viruses that subsequently integrated the *Musa* genome. Estimations of evolutionary parameters in these mixed branches would therefore be mean rates of PRVs and EPRVs. In comparison to expected rates for true EPRV branches, their evolution rate would be faster due to the rapid evolution of PRVs, and their d_N/d_S ratio would decrease due to the selective pressures acting on PRVs. Despite evidence of pseudogenization and relaxation of selective constraints, the d_N/d_S ratio for EPRVs branches was <1, a sign of functional constraints. One corollary of this hypothesis is that BSV integrations were recent since they occurred in the terminal branches of the phylogenetic tree. The second possible explanation is that several branches labeled as episomal are not true episomal viruses. In their history, they would have been endogenous for a sufficient period of time, followed by activation, i.e., an episomal step. These terminal branches would again correspond to both EPRV and PRV. In comparison to expected rates for PRV branches that did not undergo integration, their d_N/d_S ratio would increase due to the absence of selective pressures during the EPRV period, and

accordingly their evolution rate would decrease and reach the speed of neutral evolution of the *Musa* genome, i.e., at least three orders of magnitude slower than episomal viruses. Activation of EPRVs associated with the release of functional viral genomes has only been demonstrated in PVCV/Petunia, TVCV/Tobacco, and in three BSV species in banana so far; infectious EPRVs have always been considered to be rare. Our results suggest on the contrary that activation of recently integrated PRV sequences is not marginal and is much more frequent than previously thought. Endogenous badnaviruses certainly played a significant role in *Musa*–badnavirus parasitic interaction, and consequently contributed to the diversity and evolution of today's episomal BSVs.

Acknowledgments We are grateful to Nathalie Laboureau and Serge Galzi for technical assistance. We thank the members of our team, Elisabeth Fournier, Eric Bazin, and Nicolas Galtier for their helpful comments, and Philippe Rott for improving the manuscript. P.G. was supported by a PhD grant CIRAD – Région Languedoc Roussillon.

References

Ben-Dov Y (1994) A systematic catalogue of the mealybugs of the world (*Insecta*: *Homoptera*: *Coccoidea*: *Pseudococcidae* and *Putoidae*) with data on geographical distribution, host plants, biology and economic importance. Intercept Publications, Andover, p 686

Bouhida M, Lockhart BE, Olszewski NE (1993) An analysis of the complete sequence of a *Sugarcane bacilliform virus* genome infectious to banana and rice. J Gen Virol 74:15–22

Bromham L (2002) The human zoo: endogenous retroviruses in the human genome. Trends Ecol Evol 17:91–97

Daubin V, Ochman H (2004) Start-up entities in the origin of new genes. Curr Opin Genet Dev 14:616–619

Duffy S, Shackelton LA, Holmes EC (2008) Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet 9:267–276

Fauquet CM, Mayo MA, Maniloff J, Desselberger U, Ball LA (2005) Virus taxonomy, VIII report of the ICTV. Elsevier/Academic Press, London

Flint SJ, Enquist LW, Skalka AM (2003) Principles of virology: molecular biology, pathogenesis, and control of animal viruses, 2nd edn. ASM Press, Washington, DC

Gaut BS, Wright SI, Rizzon C, Dvorak J, Anderson LK (2007) Recombination: an underappreciated factor in the evolution of plant genomes. Nat Rev Genet 8:77–84

Gawel NJ, Jarret RL (1991) A modified CTAB DNA extraction procedure for *Musa* and *Ipomoea*. Plant Mol Biol Rep 9:262–266

Gayral P, Noa-Carrazana J-C, Lescot M, Lheureux F, Lockhart BEL, Matsumoto T, Piffanelli P, Iskra-Caruana M-L (2008) A single *Banana streak virus* integration event in the banana genome as the origin of infectious endogenous pararetrovirus. J Virol 82:6697–6710

Geering ADW, McMichael LA, Dietzgen RG, Thomas JE (2000) Genetic diversity among *Banana streak virus* isolates from Australia. Phytopathology 90:921–927

Geering ADW, Olszewski NE, Dahal G, Thomas JE, Lockhart BEL (2001) Analysis of the distribution and structure of integrated *Banana streak virus* DNA in a range of *Musa* cultivars. Mol Plant Pathol 2:207–213

Geering ADW, Olszewski NE, Harper G, Lockhart BEL, Hull R, Thomas JE (2005a) Banana contains a diverse array of endogenous badnaviruses. J Gen Virol 86:511–520

Geering ADW, Pooggin MM, Olszewski NE, Lockhart BEL, Thomas JE (2005b) Characterisation of *Banana streak Mysore virus* and evidence that its DNA is integrated in the B genome of cultivated *Musa*. Arch Virol 150:787–796

Gregor W, Mette MF, Staginnus C, Matzke MA, Matzke AJM (2004) A distinct endogenous pararetrovirus family in *Nicotiana tomentosiformis*, a diploid progenitor of polyploid tobacco. Plant Physiol 134:1191–1199

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696–704

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser 41:95–98

Hanada K, Suzuki Y, Gojobori T (2004) A large variation in the rates of synonymous substitution for RNA viruses and its relationship to a diversity of viral infection and transmission modes. Mol Biol Evol 21:1074–1080

Hansen CN, Harper G, Heslop-Harrison JS (2005) Characterisation of pararetrovirus-like sequences in the genome of potato (*Solanum tuberosum*). Cytogenet Genome Res 110:559–565

Harper G, Hull R (1998) Cloning and sequence analysis of *Banana streak virus* DNA. Virus Genes 17:271–278

Harper G, Osuji JO, Heslop-Harrison JSP, Hull R (1999) Integration of *Banana streak badnavirus* into the *Musa* genome: molecular and cytogenetic evidence. Virology 255:207–213

Harper G, Hart D, Moult S, Hull R (2004) *Banana streak virus* is very diverse in Uganda. Virus Res 100:51–56

Harper G, Hart D, Moult S, Hull R, Geering A, Thomas J (2005) The diversity of *Banana streak virus* isolates in Uganda. Arch Virol 150:2407–2420

Hohn T, Richert-Poggeler KR, Harper G, Schwarzacher T, Teo CH, Techeney PY, Iskra-Caruana ML, Hull R (2008) Evolution of integrated plant viruses. In: Roosinck M (ed) Virus evolution. Springer, Heidelberg, chapter 4, pp 53–81

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755

Hull R, Harper G, Lockhart B (2000) Viral sequences integrated into plant genomes. Trends Plant Sci 5:362–365

Jakowitsch J, Mette MF, van der Winden J, Matzke MA, Matzke AJM (1999) Integrated pararetroviral sequences define a unique class of dispersed repetitive DNA in plants. Proc Natl Acad Sci USA 96:13241–13246

Jaufeerally-Fakim Y, Khorugdharry A, Harper G (2006) Genetic variants of *Banana streak virus* in Mauritius. Virus Res 115:91–98

Jenkins GM, Rambaut A, Pybus OG, Holmes EC (2002) Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. J Mol Evol 54:156–165

Jones DR (2000) Diseases of banana, abacá, and enset. CABI, Wallingford

Kunii M, Kanda M, Nagano H, Uyeda I, Kishima Y, Sano Y (2004) Reconstruction of putative DNA virus from endogenous rice tungro bacilliform virus-like sequences in the rice genome: implications for integration and evolution. BMC Genomics 5:80

Le Provost G, Iskra-Caruana ML, Acina I, Teycheney PY (2006) Improved detection of episomal Banana streak viruses by multiplex immunocapture PCR. J Virol Methods 137:7–13

Lescot M, Piffanelli P, Ciampi AY, Ruiz M, Blanc G, Leebens-Mack J, da Silva FR, Santos CM, D'Hont A, Garsmeur O, Vilarinhos AD, Kanamori H, Matsumoto T, Ronning CM, Cheung F, Haas

BJ, Althoff R, Arbogast T, Hine E, Pappas GJ Jr, Sasaki T, Souza MT Jr, Miller RN, Glaszmann JC, Town CD (2008) Insights into the *Musa* genome: syntenic relationships to rice and between *Musa* species. BMC Genomics 9:58

Lheureux F, Laboureau N, Muller E, Lockhart BE, Iskra-Caruana ML (2007) Molecular characterization of *Banana streak acuminata Vietnam virus* isolated from *Musa acuminata siamea* (banana cultivar). Arch Virol 152:1409–1416

Lockhart B, Jones D (2000) Banana streak. In: Jones DR (ed) Diseases of banana, abaca and enset. CABI, Wallingford, pp 263–274

Lockhart BE, Menke J, Dahal G, Olszewski NE (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

Malik HS, Eickbush TH (2001) Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. Genome Res 11:1187–1197

Mette MF, Kanno T, Aufsatz W, Jakowitsch J, van der Winden J, Matzke MA, Matzke AJM (2002) Endogenous viral sequences and their potential contribution to heritable virus resistance in plants. EMBO J 21:461–469

Murad L, Bielawski JP, Matyasek R, Kovarik A, Nichols RA, Leitch AR, Lichtenstein CP (2004) The origin and evolution of geminivirus-related DNA sequences in *Nicotiana*. Heredity 92:352–358

Ndowora T, Dahal G, LaFleur D, Harper G, Hull R, Olszewski NE, Lockhart B (1999) Evidence that badnavirus infection in *Musa* can originate from integrated pararetroviral sequences. Virology 255:214–220

Noreen F, Akbergenov R, Hohn T, Richert-Poggeler KR (2007) Distinct expression of endogenous *Petunia vein clearing virus* and the DNA transposon dTph1 in two *Petunia hybrida* lines is correlated with differences in histone modification and siRNA production. Plant J 50:219–229

Pahalawatta V, Druffel K, Pappu H (2008) A new and distinct species in the genus Caulimovirus exists as an endogenous plant pararetroviral sequence in its host, *Dahlia variabilis*. Virology 376:253–257

Pal A, Chakrabarti A, Basak J (2007) New motifs within the NB-ARC domain of R proteins: probable mechanisms of integration of geminiviral signatures within the host species of *Fabaceae* family and implications in conferring disease resistance. J Theor Biol 246:564–573

Posada D, Buckley TR (2004) Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and bayesian approaches over likelihood ratio tests. Syst Biol 53:793–808

Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817–818

Richert-Poggeler KR, Shepherd RJ (1997) *Petunia vein-clearing virus*: a plant pararetrovirus with the core sequences for an integrase function. Virology 236:137–146

Richert-Poggeler KR, Noreen F, Schwarzacher T, Harper G, Hohn T (2003) Induction of infectious petunia vein clearing (pararetro) virus from endogenous provirus in petunia. EMBO J 22:4836–4845

Schuermann D, Molinier J, Fritsch O, Hohn B (2005) The dual nature of homologous recombination in plants. Trends Genet 21:172–181

Staginnus C, Richert-Poggeler KR (2006) Endogenous pararetroviruses: two-faced travelers in the plant genome. Trends Plant Sci 11:485–491

Staginnus C, Gregor W, Mette MF, Teo CH, Borroto-Fernandez EG, Machado ML, Matzke M, Schwarzacher T (2007) Endogenous pararetroviral sequences in tomato (*Solanum lycopersicum*) and related species. BMC Plant Biol 7:24

Su L, Gao S, Huang Y, Ji C, Wang D, Ma Y, Fang R, Chen X (2007) Complete genomic sequence of *Dracaena mottle virus*, a distinct badnavirus. Virus Genes 35:423–429

Swofford DL (2002) PAUP* 4: phylogenetic analysis using parsimony (and other methods). Sinauer, Sunderland

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Ude G, Pillay M, Nwakanma D, Tenkouano A (2002) Analysis of genetic diversity and sectional relationships in *Musa* using AFLP markers. Theor Appl Genet 104:1239–1245

Weiss RA (2006) The discovery of endogenous retroviruses. Retrovirology 3:67

Xia X (1999) DAMBE (software package for data analysis in molecular biology and evolution) user manual. Department of Ecology and Biodiversity, University of Hong Kong, Hong Kong

Xia X, Xie Z (2001) DAMBE: software package for data analysis in molecular biology and evolution. J Hered 92:371–373

Xia X, Xie Z, Salemi M, Chen L, Wang Y (2003) An index of substitution saturation and its application. Mol Phylogenet Evol 26:1–7

Yang Z (1997) PAML: a programme package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–556

Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15:568–573

Yang Z, Nielsen R (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol 46:409–418

Yang Z, Yoder AD (2003) Comparison of likelihood and Bayesian methods for estimating divergence times using multiple gene Loci and calibration points, with application to a radiation of cute-looking mouse lemur species. Syst Biol 52:705–716

Yang IC, Hafner GJ, Dale JL, Harding RM (2003) Genomic characterisation of taro bacilliform virus. Arch Virol 148:937–949

Yoder AD, Yang Z (2000) Estimation of primate speciation dates using local molecular clocks. Mol Biol Evol 17:1081–1090

Zhou Y, Holmes EC (2007) Bayesian estimates of the evolutionary rate and age of hepatitis B virus. J Mol Evol 65:197–205