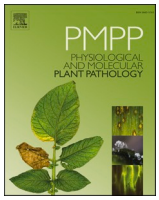
Physiological and Molecular Plant Pathology 130 (2024) 102242

Contents lists available at ScienceDirect

Physiological and Molecular Plant Pathology

journal homepage: www.elsevier.com/locate/pmpp



New insights in the evolutionary history of cacao-infecting badnaviruses

Patricia Lorena Gonzalez Grande [a,b], Fabienne Micheli [c,d,*]

[a] Centro de Computação Avançada e Multidisciplinar (CCAM), Universidade Estadual de Santa Cruz, Rodovia Ilhéus-Itabuna km16, 45662-900, Ilhéus, Bahia, Brazil
[b] Universidad Central de Venezuela, Facultad de Ciencias, Instituto de Biología Experimental, Centro de Botánica Tropical, Venezuela
[c] Centro de Biotecnologia e Genética, Universidade Estadual de Santa Cruz, Rodovia Ilhéus-Itabuna km16, 45662-900, Ilhéus, Bahia, Brazil
[d] CIRAD, UMR AGAP, F-34398, Montpellier, France

ARTICLE INFO

Keywords:
Geographic origin
Symptoms
Phylogenetics
Molecular clock
Ancestral reconstruction

ABSTRACT

The cacao crop spread throughout the world from its center of origin in South America. However, one of its main diseases cacao swollen shoot disease (CSSD), which is caused by the complex of cacao-infecting badnaviruses, was first detected in Africa. Here we investigate the relationship between the genetic diversity of the viruses that cause CSSD and their geographical distribution, and the possible relationship between the genetic diversity of the virus and the symptoms they trigger. We also sought to determine the evolutionary models that could explain this relationship and evaluate the evolutionary history of the virus through ancestral state reconstructions and temporality analyses. A high mutation rate was observed along with a clear phylogenetic signal in the geographic distribution and in the symptoms associated with each viral species. The molecular clock provided new evidence on the emergence and speciation of cacao-infecting badnaviruses, suggesting a possible American origin of these viruses.

1. Introduction

The cacao (*Theobroma cacao* L.) plant, which is of South American origin [1,2], is considered one of the main agricultural commodities [3], and its peculiar physiology restricts its cultivation to hot and humid climates, with optimal growth temperatures between 18 and 32 °C, annual rainfall of 1500–2000 mm and relative humidity of 70–80 %. Such specificity means that 75 % of commercial cacao plantations are within of a band of latitude of 8 ° on both sides of the equator, although over time it has been possible to expand its cultivation to higher latitudes (20°N and 20°S) [4,5]. Both cartographic position (latitude/longitude) and climatic conditions existing there (temperature, relative humidity and rainfall) are the main reasons for the successful introduction of cacao to the African continent during the 19th century [6].

In 1922, however, diseased cacao plants were reported in Ghanaian plantations with symptoms that had not been previously described [7]. The disease began with the appearance of red bands on the veins, followed by a foliar fern-like pattern, chlorosis and leaf mosaic. In addition to stem and root swelling, the fruits showed abnormalities, and the vigor of the plant decreased. Defoliation, foliar necrosis and death were also associated with this disease [8]. Subsequent analyses identified the *Cacao Swollen Shoot Virus* (CSSV) as the causal agent of the pathology. Characterization of CSSV showed that it is a double-stranded DNA pararetrovirus classified within the *Caulimoviridae* family and *Badnavirus* genus [9]. This genus is the most diverse of the family, and it is composed mainly of emerging pathogens from tropical areas [10].

The rapid spread of the disease to several African countries, namely, Ghana (1922), Nigeria (1944), Ivory Coast (1946) and Togo (1949) [7,11–13], motivated the study of the genetic diversity of CSSV, and high variability was detected within isolated populations [14–16]. Later, the use of next-generation sequencing showed that CSSV is not simply a virus but a complex of species that cause the so-called Cacao Swollen Shoot Disease (CSSD) [17]. Today, seven species restricted to the African continent are recognized within this viral complex: *Cacao swollen shoot CD virus* (CSSCDV), *Cacao swollen shoot CE virus* (CSSCEV), *Cacao swollen shoot Ghana M virus* (CSSGMV), *Cacao swollen shoot Ghana N virus* (CSSGNV), *Cacao swollen shoot Ghana Q virus* (CSSGQV), *Cacao swollen shoot Togo A virus* (CSSTAV), and *Cacao swollen shoot Togo B virus* (CSSTBV) [18]. There are also three different species of badnavirus that infect cacao: two American species, *Cacao mild mosaic virus* (CaMMV) and *Cacao yellow vein banding virus* (CYVBV) [19]; and one species from

Abbreviations: ML, maximum-likelihood; CSSV, Cacao Swollen Shoot Virus; CSSD, Cacao Swollen Shoot Disease.

* Corresponding author. CIRAD, UMR AGAP, Avenue Agropolis, 34398 Montpellier Cedex 5, France.
E-mail address: fabienne.micheli@cirad.fr (F. Micheli).

https://doi.org/10.1016/j.pmpp.2024.102242
Received 29 October 2023; Received in revised form 15 January 2024; Accepted 2 February 2024
Available online 3 February 2024

Asia, particularly from Sri Lanka, *Cacao bacilliform Sri Lanka virus* (CBSLV) [17].

Molecular studies of cacao-infecting badnaviruses have mainly focused on sequencing, variability analysis and recombination [14,15,17,18,20,21]. The absence of clear evidence on the association of genetic diversity and geographic distribution [8,20] as well as discrepancies and problems when trying to relate the diversity of symptoms and CSSD associated viruses genetic diversity have been reported in various studies [15,20,22,23]. Moreover, a great imbalance is observed in terms of the number of genetic sequences, studies and analysis of African badnaviruses with respect to Asian and American badnaviruses, which is probably due to the enormous impact that these viruses have had on plantations of cacao on the African continent, a scenario diametrically different from that of Asia and America [24].

In fact, more than 50% of world cacao production is currently concentrated in Ivory Coast and Ghana [3], although the devastating effect of this disease has led to large losses in both countries, which leads to an estimated reduction of 25 % in cacao plantation productivity one year after infection and up to 100 % three years later [25]. Moreover, an annual yield reduction of approximately 20,000–120,000 tons was observed in the Eastern Region of Ghana as a result of this disease [26]. The impact of these viruses negatively affects cacao plantations and severely impacts the quality of life of small farmers by considerably reducing their well-being and increasing poverty rates in households that are dedicated to this crop [27].

Consequently, despite several and repeated initiatives in the development and application of control methods for this disease, there are still no clear alternatives, moreover, issues relevant to the origin, evolutionary history, interaction processes with the plant and the evolution of the pathogenicity and virulence of the viruses are still unknown [17,28,29]. In cases like this, where a priori information is almost nil, the application of new bioinformatics tools, which combine phylogenetics and the use of metadata, besides being a widely used option, are ideal, as they allow building bridges between theoretical and empirical approaches, so that a broader and deeper approximation of the evolutionary processes studied can be achieved [30].

The “top-down” approach, for example, allows the establishment of links between a robust phylogeny and various metadata, with the aim of mapping on this new net, the points of interest and generating from it a new set of hypotheses that can be subsequently tested in an appropriate experimental setting [31–33]. This approach was chosen in the present work as a referential framework, in the search for answers to some of the previously cited questions, thus, by combining genetic data and various metadata available for cacao-infecting badnaviruses, we sought to answer the following questions: Are the geographical distribution and symptoms of the viral species associated with their molecular phylogeny? What evolutionary models could explain this phenomenon? What are the possible ancestral states of each species in relation to the geographical distribution and symptoms? What is the oldest species? All these questions are intended to contribute to drawing the evolutionary history of a virus that remains unknown in many respects despite its impact on the lives of millions of people.

2. Materials and methods

2.1. Genome sequence data

The genome sequences analyzed in the present study were downloaded from the National Center for Biotechnology Information (NCBI) database (https://www.ncbi.nlm.nih.gov/) in June 2021. In total, 86 complete cacao-infecting badnaviruses genomes (Table 1), 58 genomes of other badnaviruses (Supplementary material 1), and the *Rice Tungro Baciliform Virus* genome (ID: X57924, used as an outgroup of badnaviruses) were analyzed. The origin of all the genetic sequences of cacao-infecting badnavirus used in the present study was verified in detail, determining that according to what is described in the corresponding bibliographical references and the information provided in GenBank, all the sequences come from field isolates, therefore none of the sequences presents some type of attenuation due to continuous passages (for detailed information, see Table 1).

2.2. Genetic diversity

A pairwise nucleotide sequence comparisons were performed using Sequence Demarcation Tool (SDT) v.1.2 software [34]. For this analysis, SDT aligned the 86 full-length cacao-infecting badnavirus genome sequences, using the NW algorithms implemented in MUSCLE [35], and computed the identity score for each pair of sequences as 1-M/N, where M is the number of mismatched nucleotides and N is the total number of columns along the alignment where neither sequence has a gap character. In the next steep the program generated a rooted neighbour-joining phylogenetic tree of sequences according to which computed scores are rearranged so as to order sequences according to their likely degrees of evolutionary relatedness. Finally, SDT generated a frequency distribution of pairwise-identities using a color-coded matrix which provides intuitive accessible insights into the overall relationships between sequences.

2.3. Phylogenetic analyses

Three phylogenies were constructed in this work, and the dataset was assessed for the presence of phylogenetic signals in each phylogeny by applying the likelihood mapping analysis implemented in IQ-TREE software version 1.6.6 [36]. For the first analysis corresponding to the full cacao-infecting badnavirus phylogeny, 86 full-length cacao-infecting badnavirus genome sequences were used. For the second analysis corresponding to a partial phylogeny of cacao-infecting badnaviruses, the sequences of cacao-infecting badnaviruses of Asia (CSLBV) and America (CYVBV and CaMMV) were removed and only 80 sequences of African cacao-infecting badnaviruses were used. For both the first and second phylogenetic analyses, the sequences were aligned using the MUSCLE multiple alignment algorithm [35], with the *Commelina yellow mottle virus* (CoYMV, ID X52938) complete genome sequence used as the outgroup. The reconstruction of the tree was based on the GTR + F + I + G4 substitution model, where GTR is the generalized time reversible, F is the empirical base frequencies, I is the proportion of invariant sites, and G4 is the discrete gamma model. These models were identified through the higher Bayesian information criterion score, obtained by the ModelFinder tool [37], and implemented in the IQ-TREE software version 1.6.6 [36]. The consensus tree was built from the maximum likelihood algorithm with 1000 replicates of standard nonparametric bootstrap branch support, and the minimum bootstrap value accepted was 60. For the third analysis corresponding to the phylogeny of the cacao-infecting badnaviruses within the *Badnavirus* genus, 58 genomes of other badnaviruses (Supplementary material 1) were analyzed together with the 86 genomes of cacao-infecting badnaviruses using the *Rice Tungro Bacilliform Virus* genome (RTBV, ID: X57924) as an outgroup. A total of 145 whole genome sequences were aligned using the MAFFT online program [38]. The reconstruction of the tree was performed as described above based on the GTR + F + R7 substitution model, where R corresponds to the free rate model. Visualization of the trees was carried out on the iTOL webserver (https://itol.embl.de) [39].

2.4. Discrete traits analyzed: geographic region and symptoms

Two discrete traits were evaluated in the cacao-infecting badnavirus phylogenies and used for the tree graphic representation and model evolution design (see §2.3): Geographic region and Symptoms. The discrete trait data were obtained from the specific reference corresponding to each genome (Table 1 and Supplementary material 1). Based on the classifications proposed according to the intensity of symptom expression of stem swellings and leaf symptoms [7,40], and

Table 1
Cacao-infecting badnaviruses genomes, harvesting date and symptoms. Nd: non-determined.

Genome name/Acronym	GenBank acession number	Size (bp)	Harvesting date	Geographic region	Symptoms	Source detail	Reference
Cacao swollen shoot TB virus/CSSTBV	AJ534983	7242	1998[a]	Togo	Severe	Field isolate/ Leaves	[14]
Cacao swollen shoot TB virus/CSSTBV	AJ608931	7024	2000[a]	Ghana	Severe	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV	AJ609019	7141	2000[a]	Ghana	Mild.Swelling	Field isolate/ Leaves	
Cacao swollen shoot virus complete genome, isolate N1A	AJ609020	7006	2000[a]	Ghana	Mild.Swelling	Field isolate/ Leaves	
Cacao swollen shoot Togo A virus/CSSTAV	AJ781003	7297	2002[a]	Togo	Severe	Field isolate/ Leaves	
Cacao swollen shoot CD virus/CSSCDV	JN606110	7203	2009	Ivory Coast	Severe	Field isolate/ Leaves	[20]
Cacao mild mosaic virus isolate/CaMMV	KX276640	7533	2015	Trinidad and Tobago	Mild.Not swelling	Field isolate/ Leaves	[19]
Cacao yellow vein-banding virus/CYVBV	KX276641	7454	2015	Trinidad and Tobago	Severe.Not swelling	Field isolate/ Leaves	
Cacao swollen shoot CE virus/CSSCEV	KX592571	7020	2015	Ghana	Severe	Field isolate/ Leaves	[16]
Cacao swollen shoot CE virus/CSSCEV	KX592572	7115	2015	Ghana	Severe	Field isolate/ Leaves	
Cacao swollen shoot CE virus/CSSCEV	KX592573	7172	2012	Ivory Coast	Severe	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV[b]	KX592574	7030	2012	Ivory Coast	Nd	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV[b]	KX592575	7024	2015	Ghana	Nd	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV[b]	KX592576	6920	2012	Ivory Coast	Nd	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV[b]	KX592577	7004	2012	Ivory Coast	Nd	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV[b]	KX592578	7012	2012	Ivory Coast	Nd	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV[b]	KX592579	7025	2012	Ivory Coast	Nd	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV[b]	KX592580	7118	2012	Ivory Coast	Nd	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV[b]	KX592581	7029	2012	Ivory Coast	Nd	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV[b]	KX592582	7016	2012	Ivory Coast	Nd	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV[b]	KX592583	7022	2012	Ivory Coast	Nd	Field isolate/ Leaves	
Cacao swollen shoot CE virus/CSSCEV	KX592584	7122	2012	Ivory Coast	Severe	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV	L14546	7161	1993[c]	Togo	Severe	Field isolate	[143]
Cacao swollen shoot Togo A virus/CSSTAV[b]	MF642716	7229	2015	Ghana	Nd	Field isolate/ Leaves	[17]
Cacao swollen shoot TB virus/CSSTBV[b]	MF642717	7005	2010	Ivory Coast	Nd	Field isolate/ Leaves	
Cacao swollen shoot CD virus/CSSCDV[b]	MF642718	7205	2015	Ivory Coast	Nd	Field isolate/ Leaves	
Cacao swollen shoot CE virus/CSSCEV	MF642719	7412	2010	Ivory Coast	Severe	Field isolate/ Leaves	
Cacao swollen shoot CE virus/CSSCEV	MF642720	7131	2013	Ghana	Severe	Field isolate/ Leaves	
Cacao swollen shoot CE virus/CSSCEV	MF642721	7167	2013	Ghana	Severe	Field isolate/ Leaves	
Cacao swollen shoot CE virus/CSSCEV	MF642722	7119	2014	Ghana	Severe	Field isolate/ Leaves	
Cacao swollen shoot CE virus/CSSCEV	MF642723	6994	2014	Ghana	Severe	Field isolate/ Leaves	
Cacao swollen shoot Ghana M virus/ CSSGMV[b]	MF642724	7009	2015	Ghana	Nd	Field isolate/ Leaves	
Cacao swollen shoot Ghana N virus/ CSSGNV[b]	MF642725	7173	2015	Ghana	Nd	Field isolate/ Leaves	
Cacao swollen shoot Ghana Q virus/CSSGQV	MF642726	7091	2015	Ghana	Atypical	Field isolate/ Leaves	
Cacao swollen shoot Ghana Q virus/CSSGQV	MF642727	7343	2015	Ghana	Atypical	Field isolate/ Leaves	
Cacao swollen shoot Ghana Q virus/CSSGQV	MF642728	7102	2015	Ghana	Atypical	Field isolate/ Leaves	
Cacao swollen shoot Ghana Q virus/CSSGQV	MF642729	7186	1999	Ghana	Atypical	Field isolate/ Leaves	

(*continued on next page*)

Table 1 (*continued*)

Genome name/Acronym	GenBank acession number	Size (bp)	Harvesting date	Geographic region	Symptoms	Source detail	Reference
Cacao swollen shoot Ghana Q virus/CSSGQV	MF642730	7155	2015	Ghana	Atypical	Field isolate/ Leaves	
Cacao swollen shoot Ghana Q virus/CSSGQV	MF642731	6996	2015	Ghana	Atypical	Field isolate/ Leaves	
Cacao swollen shoot Ghana Q virus/CSSGQV	MF642732	7012	2015	Ghana	Atypical	Field isolate/ Leaves	
Cacao swollen shoot Ghana Q virus/CSSGQV	MF642733	7097	2015	Ghana	Atypical	Field isolate/ Leaves	
Cacao swollen shoot Ghana Q virus/CSSGQV	MF642734	6985	2015	Ghana	Atypical	Field isolate/ Leaves	
Cacao swollen shoot Ghana Q virus/CSSGQV	MF642735	6990	2015	Ghana	Atypical	Field isolate/ Leaves	
Cacao Bacilliform SriLanka Virus/CBSLV	MF642736	7215	2015	Sri Lanka	Mild.Swelling	Field isolate/ Leaves	
Cacao red vein-banding virus/CSSGMV	MH029281	6869	2017	Nigeria	Mild-Moderate. Swelling	Field isolate/ Leaves	[21]
Cacao red vein-banding virus/CSSGMV	MH029282	6820	2017	Nigeria	Mild-Moderate. Swelling	Field isolate/ Leaves	
Cacao red vein-banding virus/CSSGMV	MH785297	6839	2017	Nigeria	Mild-Moderate. Swelling	Field isolate/ Leaves	
Cacao red vein-banding virus/CSSGMV	MH785298	6868	2017	Nigeria	Mild-Moderate. Swelling	Field isolate/ Leaves	
Cacao red vein-banding virus/CSSGMV	MH785299	6890	2017	Nigeria	Mild-Moderate. Swelling	Field isolate/ Leaves	
Cacao red vein-banding virus/CSSGMV	MH785300	6846	2017	Nigeria	Mild-Moderate. Swelling	Field isolate/ Leaves	
Cacao red vein-banding virus/CSSGMV	MH785301	6885	2017	Nigeria	Mild-Moderate. Swelling	Field isolate/ Leaves	
Cacao red vein-banding virus/CSSGMV	MH785302	6880	2017	Nigeria	Mild-Moderate. Swelling	Field isolate/ Leaves	
Cacao red vein-banding virus/CSSGMV	MH785303	6852	2017	Nigeria	Mild-Moderate. Swelling	Field isolate/ Leaves	
Cacao swollen shoot CE virus/CSSCEV	MN433933	7314	2013	Ivory Coast	Moderate.Swelling.	Field isolate/ Leaves	[18]
Cacao swollen shoot CE virus/CSSCEV	MN433934	6978	2013	Ivory Coast	Moderate.Swelling.	Field isolate/ Leaves	
Cacao swollen shoot CD virus/CSSCDV	MN433935	7211	2016	Ivory Coast	Moderate.Swelling.	Field isolate/ Leaves	
Cacao swollen shoot CD virus/CSSCDV	MN433936	7212	2016	Ivory Coast	Moderate.Swelling.	Field isolate/ Leaves	
Cacao swollen shoot CD virus/CSSCDV	MN433937	7173	2016	Ivory Coast	Moderate.Swelling.	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV	MN433938	7090	2016	Ivory Coast	Moderate.Swelling.	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV	MN433939	7066	2016	Ivory Coast	Moderate.Swelling.	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV	MN433940	7055	2016	Ivory Coast	Moderate.Swelling.	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV	MN433941	7045	2016	Ivory Coast	Moderate.Swelling.	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV	MN433942	7045	2016	Ivory Coast	Moderate.Swelling.	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV	MN433943	7080	2016	Ivory Coast	Moderate.Swelling.	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV	MN433944	7055	2016	Ivory Coast	Moderate.Swelling.	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV	MN433945	7064	2016	Ivory Coast	Moderate.Swelling.	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV	MN433946	7063	2016	Ivory Coast	Moderate.Swelling.	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV	MN433947	7007	2016	Ivory Coast	Moderate.Swelling.	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV	MN433948	7027	2016	Ivory Coast	Moderate.Swelling.	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV	MN433949	7027	2016	Ivory Coast	Moderate.Swelling.	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV	MN433950	7031	2016	Ivory Coast	Moderate.Swelling.	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV	MN433951	7024	2016	Ivory Coast	Moderate.Swelling.	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV	MN433952	7018	2016	Ivory Coast	Moderate.Swelling.	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV	MN433953	7023	2016	Ivory Coast	Moderate.Swelling.	Field isolate/ Leaves	

(*continued on next page*)

Table 1 (*continued*)

Genome name/Acronym	GenBank acession number	Size (bp)	Harvesting date	Geographic region	Symptoms	Source detail	Reference
Cacao swollen shoot TB virus/CSSTBV	MN433954	7041	2016	Ivory Coast	Moderate.Swelling.	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV	MN433955	7023	2016	Ivory Coast	Moderate.Swelling.	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV	MN433956	7040	2016	Ivory Coast	Moderate.Swelling.	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV	MN433957	7030	2016	Ivory Coast	Moderate.Swelling.	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV	MN433958	7035	2016	Ivory Coast	Moderate.Swelling.	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV	MN433959	7022	2016	Ivory Coast	Moderate.Swelling.	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV	MN433960	7018	2016	Ivory Coast	Moderate.Swelling.	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV	MN433961	7024	2016	Ghana	Severe	Field isolate/ Leaves	
Cacao swollen shoot TB virus/CSSTBV	MN433962	7072	2016	Ghana	Severe	Field isolate/ Leaves	
Cacao mild mosaic virus/CaMMV	MW052520	7514	2019	Puerto Rico	Mild.Not swelling	Field isolate/ Leaves	[105]
Cacao mild mosaic virus/CaMMV	MW052521	7520	2019	Brazil	Mild.Not swelling	Field isolate/ Leaves	
Cacao mild mosaic virus/CaMMV	MW052522	7524	2019	Brazil	Mild.Not swelling	Field isolate/ Leaves	

[a] Data obtained through personal communication with the authors of the corresponding cited reference.
[b] Genome excluded from the analysis of ancestral character reconstruction and evolutionary model analysis.
[c] Genome without exact harvesting date; the date of sequence submission in GenBank was used as harvesting date.

the set of symptoms described for all genomes evaluated (referenced in Table 1), the data were sorted into 7 symptom categories: Mild.Swelling, Mild-Moderate.Swelling, Moderate.Swelling, Severe, Atypical, Mild.Not swelling, Severe.Not swelling (Supplementary material 2).

2.5. Evolution models and phylogenetic signals

The mode of evolution of the two discrete traits was analyzed in *R* version 4.0.3 [41] using the fitDiscrete function from the Geiger package [42]. This function uses likelihood analyses to test how well the observed discrete data and phylogeny fit models of evolution. Five models were tested for each trait: i) equal rates (ER) model, where transitions may occur between all possible combinations of states in any direction, and the rates of transition are all equal; ii) all-rates-different (ARD) model, where each rate is a unique parameter; iii) symmetric model (SYM), where forwards and reverse transitions share the same parameter; iv) Meristic standard model, where transitions occur in a stepwise fashion and only occur between adjacent states and with an equal rate; and v) all-rates-different Meristic model (Meristic_ARD), where the transition occurs in an ordered mode but where the transition rate into adjacent states can be different and not necessarily equal. For the last model, which was not available in the default models of the fitDiscrete function, we developed a specific script containing a matrix representing the transition classes between states (see *R* code in Supplementary material 3). After model analysis, different tree transformation types were tested: i) lambda type [43,44], which corresponds to a common measure of the phylogenetic signal and fits the extent to which the phylogeny predicts covariance among trait values for tips; default bounds for lambda model are min = 0, max = 1, high values representing high phylogenetic signal; ii) kappa type, which corresponds to a punctuational (speciation) model of trait evolution [45] and includes default bounds of min = 0, max = 1; iii) delta type that is a time-dependent model of trait evolution [43,44], with default bounds of min = 0, max = 3, where if delta is greater than 1, the recent evolution has been relatively fast, and if delta is lesser than 1, the recent evolution has been comparatively slow; and iv) the early-burst (EB) type [42], which is where the rate of evolution increases or decreases exponentially through time, with default bounds of min = −10, max = 10. The corrected Akaike information criterion (AICc) and the log likelihood, both supplied by the fitDiscrete function, as well as the delta AIC (ΔAIC) and the AIC weight (AICw), both supplied by the package Geiger function aicw, were used to determine the models that best fit the studied data (Supplementary material 3). The AICc was used instead of the AIC because it is recommended for small sample sizes and is an overall more robust comparative metric [46]. In the specific case of symptoms for which some data were missing (genomes without information; Table 1), phylogenetic tree tips were dropped using the function drop.tip from the ape package version 5.4-1 [47] in *R* version 4.0.3 [41]. According to the results of phylogenetic analysis and considering possible bias related to disequilibrium in sample amount (80 African genome sequences *vs.* 6 non-African genome sequences; Supplementary materials 4 and 5) for evolution model establishment [48], we focused the evolution analysis on the 80 sequences of the African cacao-infecting badnaviruses.

2.6. Ancestral state reconstruction

The reconstruction of the ancestral state of Geographic region and Symptoms was performed using the two maximum-likelihood (ML) cacao-infecting badnavirus phylogenies described above (§2.3.), through the PastMl software v.1.9.30 [49]. PastML uses decision-theory concepts to associate each node in the tree with a set of likely states. In the easy regions of the tree (typically close to the tips), the program predicts a unique state, whereas in the more challenging parts of the tree, it may predict several likely states, thus reflecting the uncertainty of the inferences. The options for running PastML in the current study were the prediction method, marginal posterior probabilities approximation (MPPA) and Model Felsenstein 1981 (F81), which provide the most accurate ancestral predictions [49]. The software provided two graphical outputs: a large phylogeny with the ancestral states for each node and a compressed representation of the ancestral scenarios where only the main information was highlighted. Both outputs are presented in the results section. As described above (§2.4), we focused on the reconstruction analysis of the 80 sequences of the African cacao-infecting badnaviruses (analysis with all 86 full-length cacao-infecting badnaviruses genomes is shown in Supplementary materials 6 and 7).

2.7. Temporal signal and molecular clock

To investigate the temporal signal, we regressed root-to-tip genetic distances from the full cacao badnavirus maximum-likelihood (ML) phylogeny described above (§2.3) against sample harvesting dates (Table 1) using TempEst v.1.5.1 [50]. Here, the dataset was annotated with sampling years between 1990 and 2019, except for genome L14546, which did not have an exact collection date; in this last case, the submission date of the sequence in GenBank was taken as the collection date (Table 1). According to Navascués et al. [51], the significance of the regression was assessed by random permutation of the sampling dates over the sequences, with the correlation coefficient used as the test statistic. Subsequently, 1000 replicates of the data were generated, with the sampling dates randomly permuted, using the approach of Murray et al. [52]. The p value of the root-to-tip regression was calculated using the function rootToTipRegressionPlot from the package Treedater [53] in *R* v.4.0.3 [41]. Furthermore, prior to fitting a dated-tip molecular clock, we evaluated the possible problems of confounding temporal and genetic structures using the Mantel confounding test developed in *R* by Murray et al. [52]. In the absence of confounders, the molecular clock analysis was carried out in MEGA X [54] using the RelTime method [55, 56] and the sampling dates (Table 1). Visualization of the tree was performed using the iTOL webserver (https://itol.embl.de) [39].

3. Results

3.1. Genetic diversity in cacao-infecting badnaviruses

Differences in relation to genetic diversity were observed within each of the CSSD associated species (Supplementary material 8). CSSGQV contained two well-defined internal groups, one with a high percentage of identity (90 %) and the other with less identity (70 %), and the identity of the total group was 85.36 %. CSSGQV showed 60 % identity with its closest neighbour CSSCEV. CSSTAV was represented by two genomes that shared an identity greater than 86.94 %; with respect to their closest neighbours CSSGQV and CSSCEV, the identity was approximately 60 and 70 %, respectively. CSSCDV was a homogeneous group with an internal identity greater than 90 %, and it shared 60, 70 and 77 % identity with CSSGQV, CSSCEV and CSSTAV, respectively. CSSTBV was the group with the highest number of genomes; however, it was a genetically very homogeneous group (93.3 % identity) that shared 60, 70, 78 and 77 % identity with CSSGQV, CSSCEV, CSSTAV and CSSCDV, respectively. The isolate N1A (GenBank accession number AJ609020), which was not previously classified, was grouped within the CSSTBV species. CSSGMV, as CSSGQV, has two internal groups that shared 91.64 % identity, which was higher than that found within the CSSGQV species. CSSGMV shared 60 % identity with CSSGQV, 70 % identity with CSSCEV and 75 % identity with CSSTAV, CSSCDV and CSSTBV. CSSGNV, which was represented by a single genome, shared 60 % identity with CSSGQV, 70 % identity with CSSCEV and 75 % identity with CSSTAV, CSSCDV, CSSTBV and CSSGMV. CSSCEV was the least homogeneous group (78.26 %). When ordering the African CSSD associated species according to the pairwise sequence comparisons (percentage of genetic identity of the complete genome), the following order was obtained from highest to lowest identity: CSSTBV, CSSCDV, CSSGMV, CSSTAV, CSSGQV and CSSCEV. Finally, CaMMV presented two well-differentiated groups with 93.75% identity, while CBSLV, CaMMV and CYVBV were distant from the badnaviruses that infected cacao in Africa, sharing only 60 % genetic identity with them (Supplementary material 7).

3.2. Likelihood mapping

The phylogenetic signal from each dataset was investigated by likelihood-mapping analysis [36]. Three datasets were evaluated: the complete genomes of all cacao-infecting badnaviruses (86 sequences), the complete genomes of African cacao-infecting badnaviruses (80 sequences), and the complete genomes of badnaviruses (144 sequences). The likelihood-mapping analysis for the cacao-infecting badnaviruses revealed that the quartets from the dataset were primarily distributed in the corners (92.6 %) rather than the sides (3 %) or center (4.3 %) of the triangle. For the second and third datasets, the distribution patterns were similar: 91.4 % in the corners, 3.3 % in the sides and 5.2 % in the center and 95.8 % in the corners, 2.8 % in the sides and 1.4 % in the center, respectively (Supplementary material 9). These results coincided with the premise that robust phylogenetic signals translate into a high number of resolved quartets and a low number of unresolved quartets. Therefore, the likelihood mappings of the evaluated sequences indicate that they have a strong tree-like phylogenetic signal and are suitable for phylogenetic reconstruction.

3.3. Cacao-infecting badnavirus phylogeny in the context of geographic region and symptoms

To evaluate the quality of the phylogenetic signal of the studied traits (Geographic region and Symptoms) throughout the molecular phylogeny built with the complete genomes of cacao-infecting badnaviruses, Pagel's lambda test was performed [43,44]. Pagel's lambda is a measure of the phylogenetic signal, and its degree represents the influence of the phylogenetic history on the distribution of the traits observed at the tips. As observed in the tree, the two evaluated traits Geographic region ($\lambda = 1$) and Symptoms ($\lambda = 1$) presented high values of lambda, indicating a strong phylogenetic signal between the given traits and the built phylogeny. Non-African badnavirus species (CBSLV, CYVBV and CaMMV) that were present at the base of the phylogenetic tree were genetically distant from African species (Fig. 1). However, the CaMMV species had a different ancestor to CYVBV, although the two species are from Trinidad and Tobago in America. In fact, CaMMV shared a common ancestor much closer to Africa's badnaviruses than CYVBV, which appeared in another group alongside CBSLV (Fig. 1). The badnavirus identified in Sri Lanka (CBSLV) generated mild symptoms in cacao, whereas the badnaviruses CaMMV and CYVBV identified in America had the capacity to generate mild and moderate symptoms, respectively, and did not generate the typical swelling identified in Africa (Fig. 1, Table 1). The most basal cacao-infecting African badnavirus species was CSSGQV (Fig. 1), and it was subdivided into two genetically well differentiated groups. However, all available genomes of this species share the same Geographic region (Ghana) and generated the same Symptoms (atypical, they were only able to coinfect, they did not infect alone; Fig. 1, Table 1). The CSSGNV species, for which only one genome and no symptom information were available (Table 1), originated from Ghana and appeared at the base of the branch containing the CSSGMV species, which seems to indicate a genetic proximity between these two species (Fig. 1). The CSSGMV species showed two genetically differentiated subgroups, and it was mostly present in Nigeria (only one genome was identified in Ghana) and the plant symptoms caused by all members of this species were mild-moderate with swelling (Fig. 1, Table 1). CSSCEV was identified in Ghana and Ivory Coast and formed a phylogenetically compact group that showed severe symptoms for most of the available genomes (Fig. 1, Table 1); moreover, this species was associated with rapid death syndrome in cacao trees. The CSSTAV isolates that also corresponded to a genetically homogeneous group were collected in Ghana and Togo, with the Togo isolates generating severe symptoms in cacao trees (Fig. 1, Table 1). The CSSCDV isolates were phylogenetically close to the CSSTAV isolates and previously identified in the Ivory Coast, and most of them (except one isolate that generates severe symptoms) led to moderate swelling symptoms (Fig. 1, Table 1). At the upper end of the phylogenetic tree, and farther from the ancestors, CSSTBV species were found, and they were identified in three countries: Ghana, Ivory Coast and Togo (Fig. 1, Table 1). Isolates originating from Togo had the ability to generate severe symptoms in cacao plants; isolates from the Ivory Coast were less pathogenic, generating moderate symptoms with

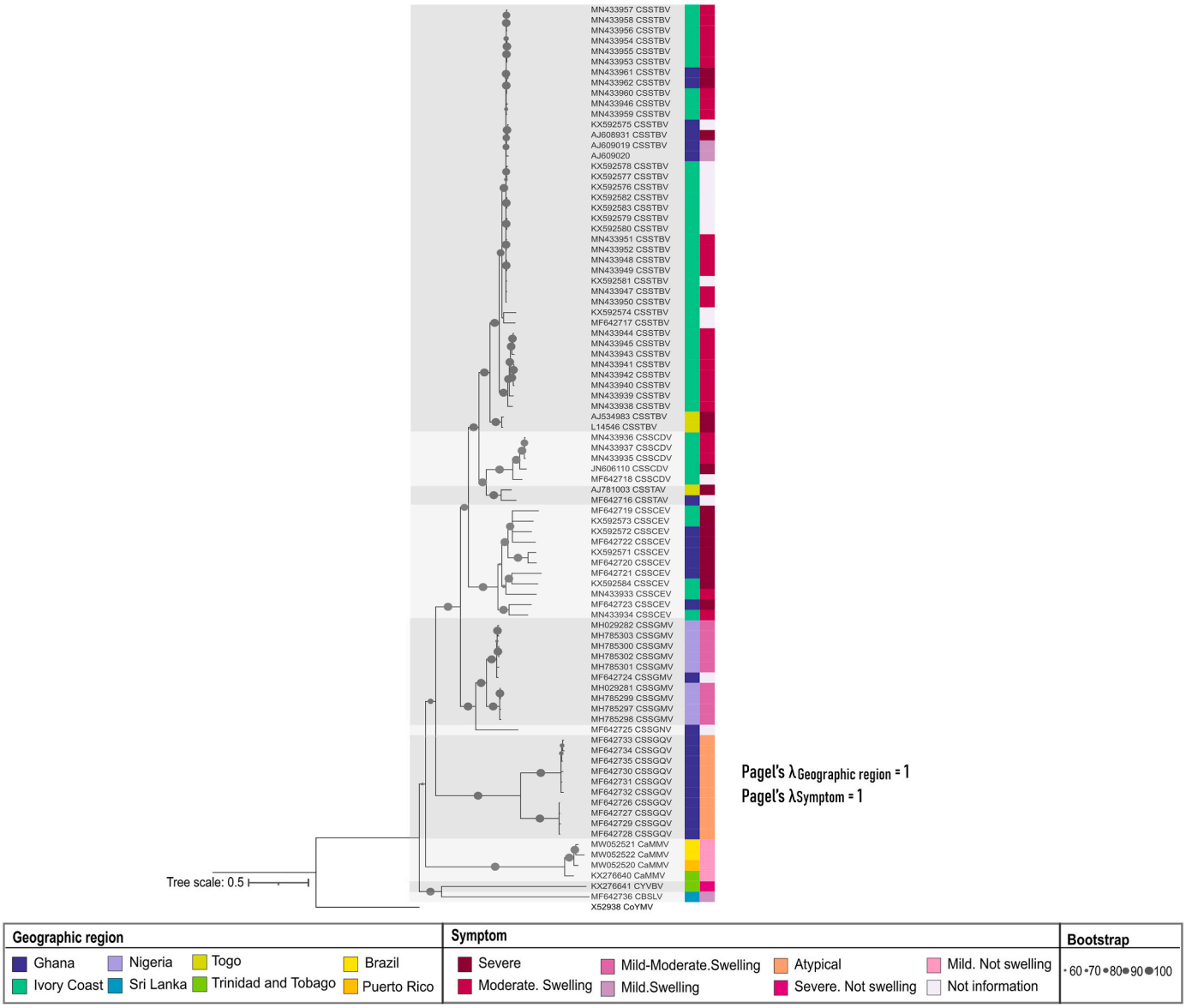


Fig. 1. Maximum likelihood phylogenetic tree based on alignment of the complete genomes of 86 cacao-infecting badnaviruses. The tree reconstruction was obtained using the GTR + F + I + G4 substitution model with 1000 bootstrap repetitions. Grey circles represent the bootstrap values for each branch (from 60 to 100). The two columns on the right of the figure represents the Geographic region and the Symptoms (left and right columns, respectively; color legend at the bottom of the figure). The accession number precedes the species abbreviation: CBSLV, *Cacao bacilliform Sri Lanka virus*; CaMMV, *Cacao mild mosaic virus*; CYVBV, *Cacao yellow vein banding virus*; CSSCDV, Cacao swollen shoot CD virus; CSSCEV, *Cacao swollen shoot CE virus*; CSSGMV, *Cacao swollen shoot Ghana M virus*; CSSGNV, *Cacao swollen shoot Ghana N virus*; CSSGQV, *Cacao swollen shoot Ghana Q virus*, CSSTAV, *Cacao swollen shoot Togo A virus*, and CSSTBV, *Cacao swollen shoot Togo B virus*. *Commelina Yellow Mottle Virus* (CoYMV) was used as an outgroup.

swelling; and isolates from Ghana generated both severe symptoms and mild symptoms with swelling (Fig. 1, Table 1). As observed in the distance matrix (Supplementary Material 8), the N1A isolate, without previous identification, was robustly grouped within the CSSTBV species, suggesting that N1A could be a member of this species. Ghana had the greatest diversity of CSSD associated species with six different species, followed by Ivory Coast with three, Togo with two and Nigeria with a single species (Table 1, Fig. 1, Supplementary material 10). In America and Asia, one species of cacao-infecting badnavirus was recorded per country with the exception of Trinidad and Tobago, where two species have been recorded (Table 1, Fig. 1). Ghana is also the country that presented the greatest diversity of Symptoms, which may be due to the diversity of species existing in the territory. In relation to the cacao-infecting badnavirus species, those with the greatest distribution were CSSTBV (present in three African countries) and CaMMV (present in three American countries), followed by CSSCEV, CSSGMV and CSSTAV (each present in two African countries). The remaining species were registered in a single country (Fig. 1).

3.4. Evolution models for geographic region and symptoms of cacao-infecting badnaviruses

Regarding the Geographic region trait for African cacao-infecting badnaviruses, the evolution model that presented the best overall fit, i.e., the lowest AICc and ΔAIC and highest AICw were symmetric model with EB transformation (SYM_EB model) (Table 2, Fig. 2A). In this model, forward and reverse transitions between the trait states shared the same parameter and occurred exponentially over time because the EB transformation was associated with exponential changes (increases or decreases) in the evaluated trait over time. According to this model,

Table 2
Results of likelihood models and transformation fitting analyses of Geographic region and Symptoms over complete genome phylogenies of cacao-infecting African badnaviruses. Models: ARD: All Rates Different; ER: Equal-Rates; SYM: Symmetric; Meristic: Standard; Meristic_ARD: All Rates Different Meristic. Transformation: Delta, Kappa, Lambda; EB: Early-burst model. Df: Degrees of freedom. AICc: Corrected Akaike Information Criterion. ΔAIC: Delta AIC. AICw: AIC weight. Best fitting model is highlighted in grey.

Trait	Model / Transformation	df	log-likelihood	AIC			Parameter value
				AICc	ΔAIC	AICw	
Geographic region	ER	1	-50.155	102.362	3.27E+00	8.64E-02	-
	ARD	12	-50.155	108.703	9.62E+00	3.63E-03	-
	SYM	6	-49.329	104.222	5.13E+00	3.41E-02	-
	Meristic	3	-49.640	110.097	1.10E+01	1.81E-03	-
	Meristic_ARD	12	-49.254	108.703	9.62E+00	3.63E-03	-
	ER_ Lambda	2	-40.023	104.466	5.38E+00	3.02E-02	1
	ER_Kappa	2	-40.023	102.813	3.73E+00	6.90E-02	0.635871
	ER_Delta	2	-38.979	103.437	4.35E+00	5.05E-02	0.006738
	ER_EB	2	-37.550	102.664	3.58E+00	7.43E-02	-4.14282
	ARD_Lambda	13	-36.705	111.561	1.25E+01	8.69E-04	1
	ARD_Kappa	13	-45.536	109.473	1.04E+01	2.47E-03	0.681943
	ARD_Delta	13	-45.536	106.615	7.53E+00	1.03E-02	2.999999
	ARD_EB	13	-43.764	104.925	5.84E+00	2.40E-02	6.727492
	SYM_ Lambda	7	-44.636	106.627	7.54E+00	1.02E-02	1
	SYM_Kappa	7	-41.766	103.084	4.00E+00	6.02E-02	0.546664
	SYM_Delta	7	-51.890	104.828	5.74E+00	2.52E-02	0.006738
	SYM_EB	7	-51.890	99.087	0.00E+00	4.44E-01	-7.784358
	Meristic_ Lambda	4	-50.936	112.314	1.32E+01	5.96E-04	1
	Meristic_Kappa	4	-51.206	110.406	1.13E+01	1.55E-03	0.716159
	Meristic_Delta	4	-48.045	110.946	1.19E+01	1.18E-03	0.006738
	Meristic_EB	4	-40.023	104.623	5.54E+00	2.79E-02	-7.334638
	Meristic_ARD_ Lambda	13	-40.023	111.561	1.25E+01	8.69E-04	1
	Meristic_ARD_Kappa	13	-38.979	109.473	1.04E+01	2.47E-03	0.68194
	Meristic_ARD_Delta	13	-37.550	106.615	7.53E+00	1.03E-02	2.999999
	Meristic_ARD_EB	13	-36.705	104.925	5.84E+00	2.40E-02	6.728158
Symptoms	ER	1	-46.096	94.256	8.96E+00	4.56E-03	-
	ARD	20	-46.096	121.244	3.59E+01	6.29E-09	-
	SYM	10	-41.964	99.999	1.47E+01	2.58E-04	-
	Meristic	4	-45.686	88.647	3.35E+00	7.54E-02	-
	Meristic_ARD	20	-45.294	121.244	3.59E+01	6.29E-09	-
	ER_ Lambda	2	-31.077	96.386	1.11E+01	1.57E-03	1
	ER_Kappa	2	-31.077	88.122	2.83E+00	9.80E-02	0.35984
	ER_Delta	2	-30.336	95.565	1.03E+01	2.37E-03	0.006738
	ER_EB	2	-30.954	94.781	9.49E+00	3.51E-03	-3.992553
	ARD_Lambda	21	-30.351	125.642	4.03E+01	6.98E-10	1
	ARD_Kappa	21	-37.962	124.160	3.89E+01	1.46E-09	0.578325
	ARD_Delta	21	-37.357	125.397	4.01E+01	7.89E-10	1.865971
	ARD_EB	21	-35.440	124.191	3.89E+01	1.44E-09	-5.715725
	SYM_ Lambda	11	-36.416	101.695	1.64E+01	1.11E-04	0.953592
	SYM_Kappa	11	-34.628	97.861	1.26E+01	7.52E-04	0.406031
	SYM_Delta	11	-39.990	99.812	1.45E+01	2.84E-04	0.006738
	SYM_EB	11	-39.990	96.238	1.09E+01	1.69E-03	-7.946074
	Meristic_ Lambda	5	-37.395	90.997	5.70E+00	2.33E-02	1
	Meristic_Kappa	5	-38.840	85.807	5.12E-01	3.12E-01	0.397547
	Meristic_Delta	5	-37.139	88.698	3.40E+00	7.35E-02	0.019973
	Meristic_EB	5	-31.077	85.295	0.00E+00	4.03E-01	-7.697864
	Meristic_ARD_ Lambda	13	-40.023	111.561	2.63E+01	7.97E-07	1
	Meristic_ARD_Kappa	13	-38.979	109.473	2.42E+01	2.26E-06	0.68194
	Meristic_ARD_Delta	13	-37.550	106.615	2.13E+01	9.45E-06	2.999999
	Meristic_ARD_EB	13	-36.705	104.925	1.96E+01	2.20E-05	6.728158

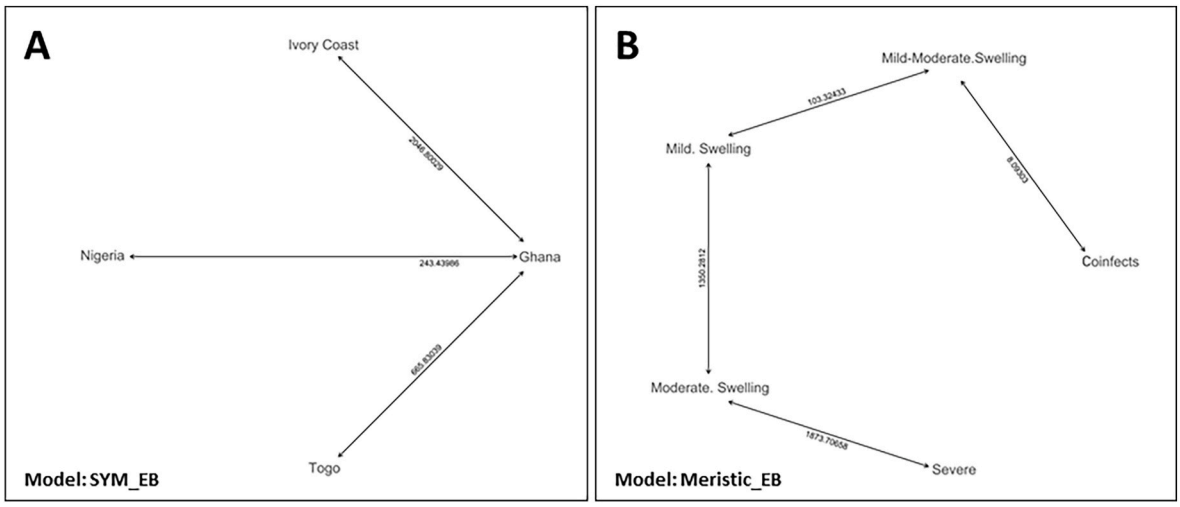


Fig. 2. Schemes of the transitions between trait states reconstructed for cacao-infecting African badnaviruses. **A**. Best fitted evolution model for the Geographic region trait. **B**. Best fitted evolution model for the Symptoms trait. Models: SYM: symmetric; Meristic: standard. Transformation: EB: Early-burst model.

the main source of the virus in Africa was Ghana, and from there, the virus spread to Nigeria, Togo and the Ivory Coast (Fig. 2A). The Symptoms differences between the African cacao-infecting badnaviruses better fit the Meristic_EB model (Table 2, Fig. 2B). This model is characterized by the fact that transitions occur in an orderly manner, and as in the previous case, the transitions occur exponentially over time (EB

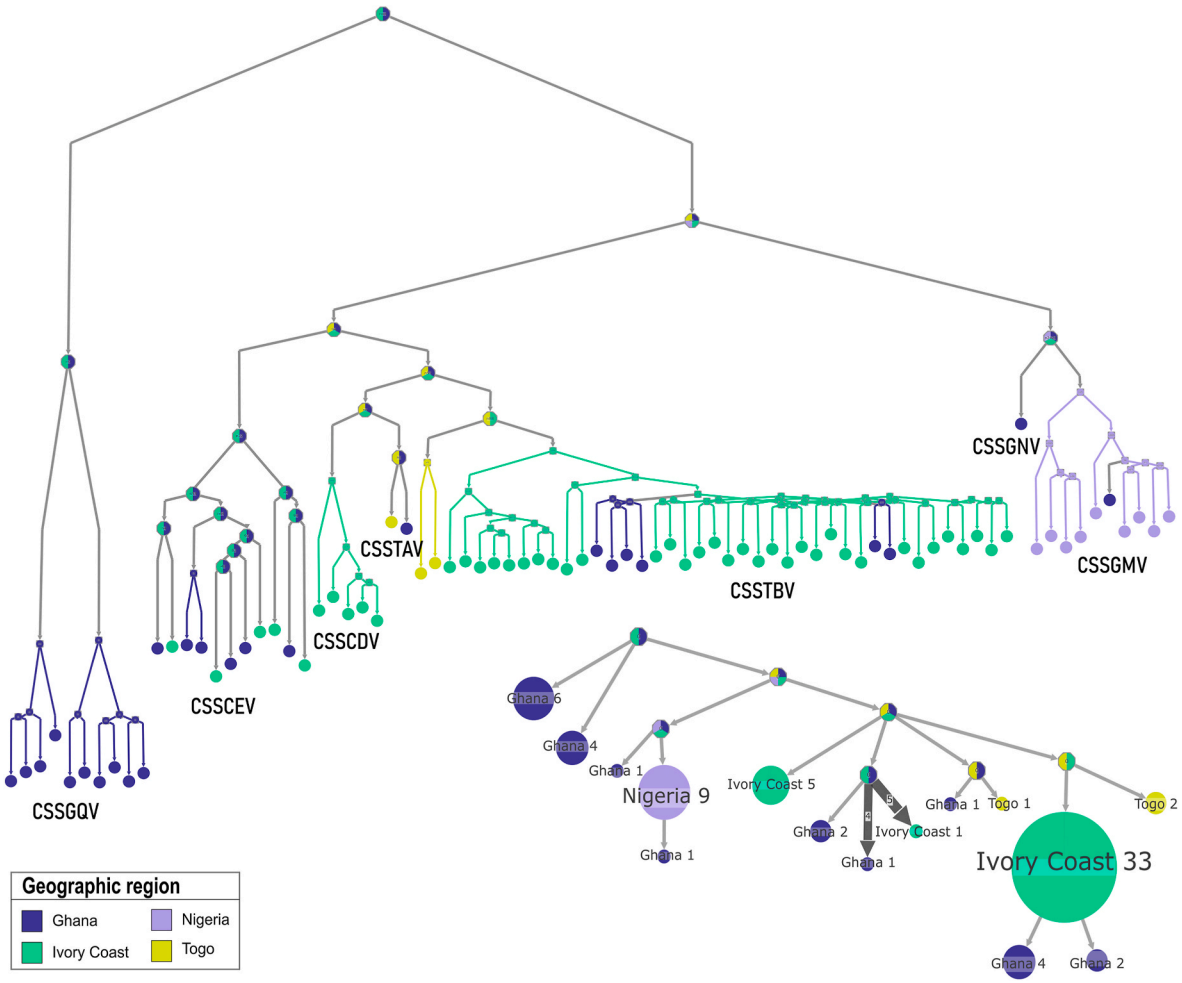


Fig. 3. Ancestral state reconstruction of Geographic region of cacao-infecting African badnaviruses. The tree and compressed visualizations were obtained using PastML with MPPA + F81 option (see §2.6). Different colors correspond to different geographical regions as shown in the legend in the bottom left corner. CSSCDV, *Cacao swollen shoot CD virus*; CSSCEV, *Cacao swollen shoot CE virus*; CSSGMV, *Cacao swollen shoot Ghana M virus*; CSSGNV, *Cacao swollen shoot Ghana N virus*; CSSGQV, *Cacao swollen shoot Ghana Q virus*; CSSTAV, *Cacao swollen shoot Togo A virus*; CSSTBV, *Cacao swollen shoot Togo B virus*.

transformation). According to the model with the best fit, a possible evolution of Symptoms of CSSD associated viruses would start from viruses that were only capable of coinfection, and then mild-moderate symptoms would appear, followed by moderate and severe symptoms (Fig. 2B).

3.5. Ancestral state reconstruction of geographic region and symptoms in cacao-infecting badnaviruses

The reconstruction of the ancestral states of Geographic region allowed us to identify as possible regions of cacao-infecting badnavirus origin the countries Ghana and/or Ivory Coast (Fig. 3). From one of these regions, a possible early dispersal event towards Nigeria and/or Togo could have occurred, which may have generated an ancestor that gave rise to the endemic species of Nigeria (CSSGMV). On the other hand, the analysis suggested that an ancestor from Togo and/or Ivory Coast could have given rise to the CSSTVB species distributed in Ghana, Ivory Coast and Togo (Fig. 3). While the most recent ancestor of the CSSCDV species could be located in Ghana, Ivory Coast or Togo, the most recent ancestor of CSSTAV was limited to Togo or Ghana. The most recent ancestor of CSSCEV was clearly located in Ghana or the Ivory Coast; these countries are also possible CSSGQV centers of origin (Fig. 3). Despite the great imbalance in the number of sequences for each geographic region, the reconstruction of ancestral states for all the available genomes of cacao-infecting badnaviruses was performed (Supplementary material 6). In the case of CaMMV, its most recent ancestor could have come from Brazil (Supplementary material 6). Regarding the Symptoms trait for African cacao-infecting badnaviruses, the reconstruction carried out suggested a first ancestor causing Severe or Moderate.Swelling symptoms that would give rise to two intermediate ancestors (Fig. 4). One of them caused Severe or Atypical symptoms and would give rise to the CSSGQV species, and they all present atypical pathogenicity since they only coinfect. A second intermediate ancestor causing medium to severe symptoms (Mild-Moderate Swelling, Moderate.Swelling or Severe) that could have originated the CSSGMV species (causing Mild-Moderate. Swelling symptoms) and an intermediate ancestor causing Moderate. Swelling or Severe symptoms, which would give origin to could have given rise to CSSCEV, CSSTAV, CSSCDV and CSSTBV (Fig. 4).

3.6. Cacao-infecting badnaviruses in the badnavirus context

The phylogenetic relationships between cacao-infecting badnaviruses and the available genomes of other badnaviruses showed that CBSLV and CYVBV presented a large distance from the cacao-infecting badnavirus pool (Fig. 5). CAMMV, although appearing distant, was considerably closer to the other cacao-infecting badnaviruses (Fig. 5), as also observed in the phylogenetic analysis (Fig. 1). Interestingly, the CSSGQV species was completely separated from the other cacao-infecting badnavirus species and clustered with *Mulberry badnavirus-1* (MBV-1) (Fig. 5). Other cacao-infecting badnavirus species clustered

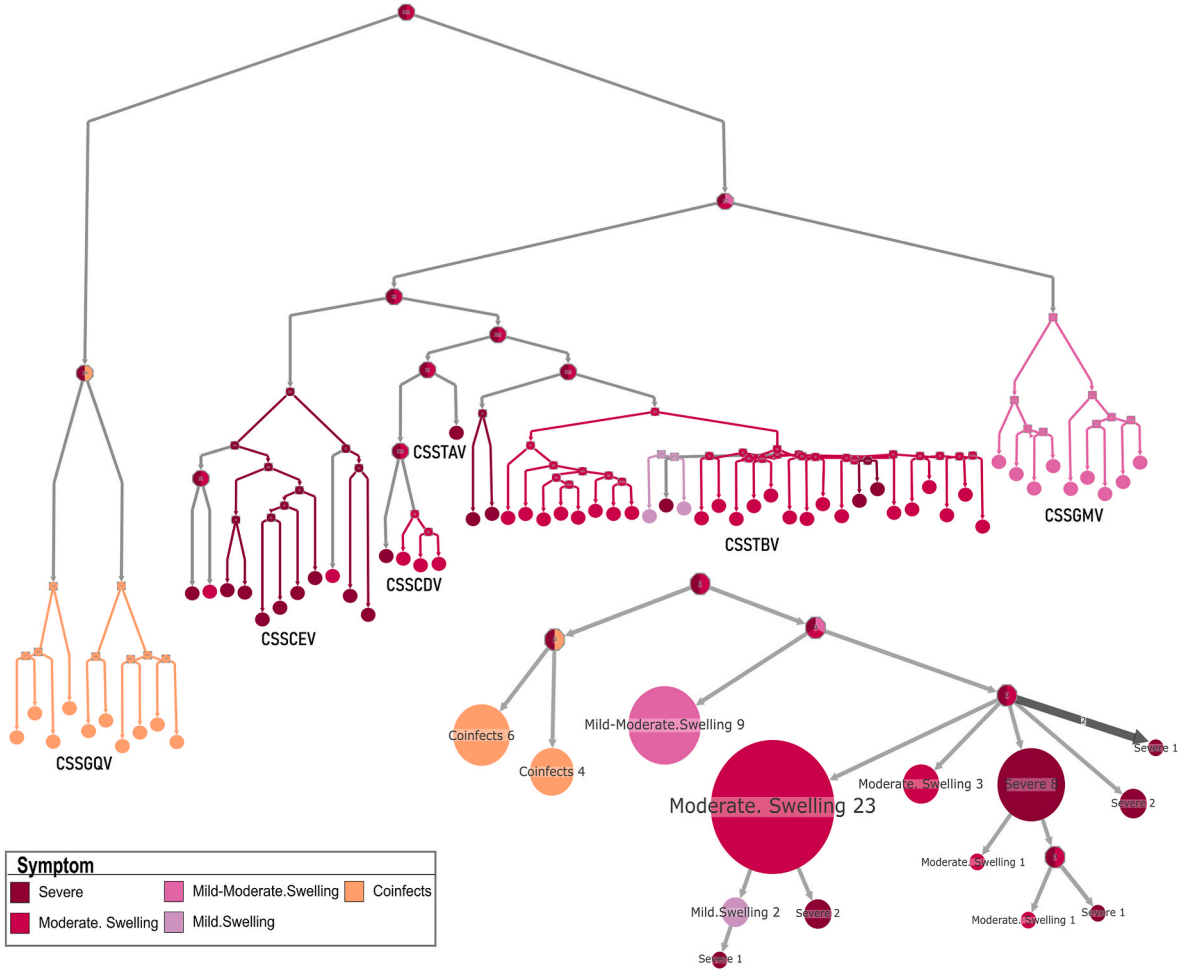


Fig. 4. Ancestral state reconstruction of Symptoms of cacao-infecting African badnaviruses. The tree and compressed visualizations were obtained using PastML with MPPA + F81 option (see §2.6). Different colors correspond to different symptoms as shown in the legend in the bottom left corner. CSSCDV, *Cacao swollen shoot CD virus*; CSSCEV, *Cacao swollen shoot CE virus*; CSSGMV, *Cacao swollen shoot Ghana M virus*; CSSGQV, *Cacao swollen shoot Ghana Q virus*; CSSTAV, *Cacao swollen shoot Togo A virus*; CSSTBV, *Cacao swollen shoot Togo B virus*.

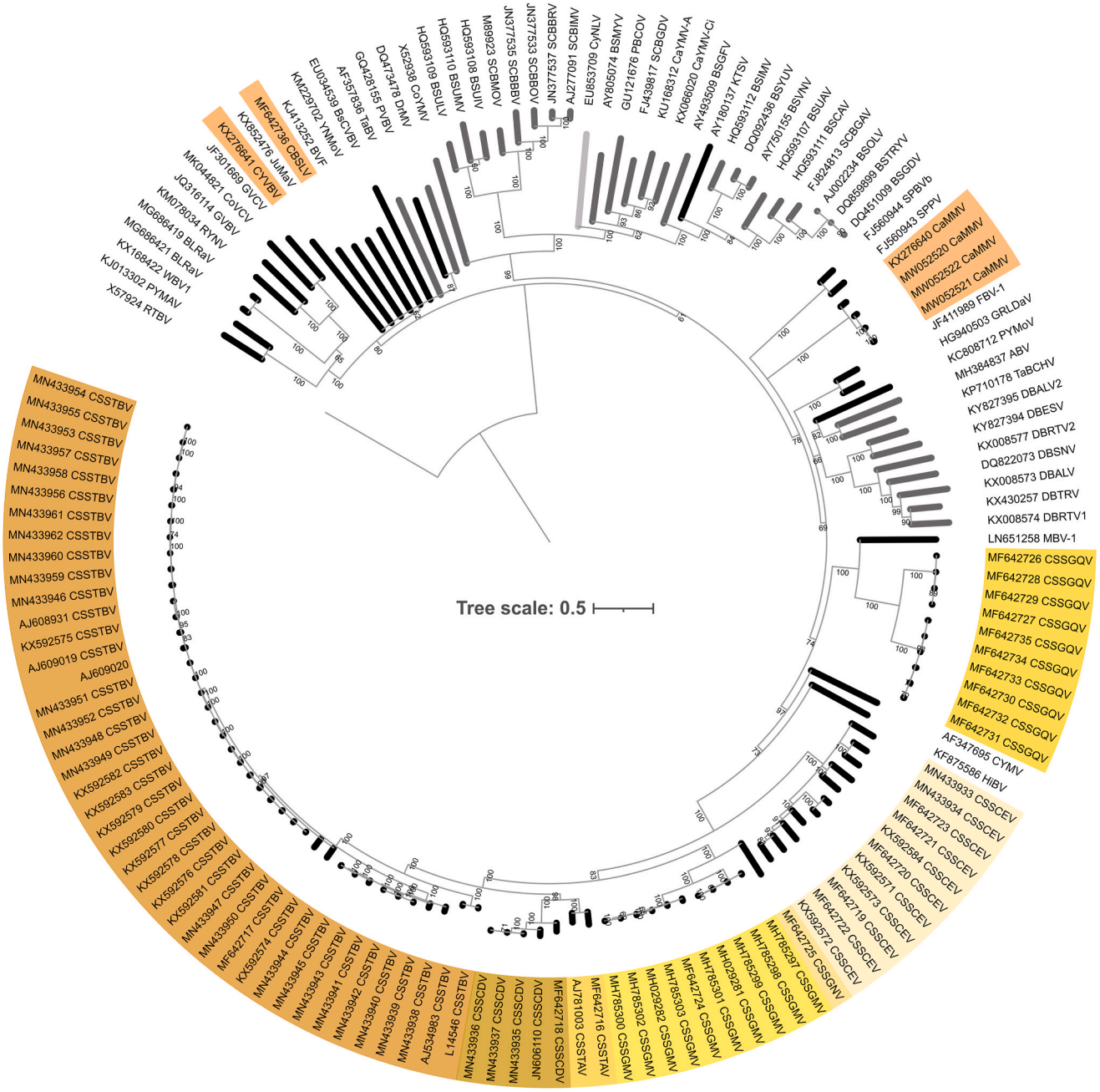


Fig. 5. Maximum likelihood phylogenetic tree based on alignment of the complete genomes of 144 badnaviruses. Model Substitution GTR + F + R6 and 1000 replicated bootstrap analysis. The bootstrap values are indicated in each branch. Minimum bootstrap value 60. Each tip shows the GenBank accession number and the species name. The color of branch indicates the Class of plants infected by this virus (grey branch corresponds to Liliopsida and black branch to Magnoliopsida). The species highlights in colors correspond to the cacao-infecting badnaviruses. ABV, *Aglaonema bacilliform virus*; BSCAV, *Banana streak CA virus*, BSGDV, *Banana streak GD virus*; BSGFV, *Banana streak GF virus*; BSIMV: *Banana streak IM virus*; BSMYV, *Banana streak Mysore virus*; BSOLV, *Banana streak OL virus*; BSTRYV, *Banana streak TRY virus*; BSUAV, *Banana streak UA virus*; BSUIV, *Banana streak UI virus*; BSULV, *Banana streak UL virus*; BSUMV, *Banana streak UM virus*; BSYUV, *Banana streak virus Acuminata Yunnan*; BSVNV, *Banana streak virus strain Acuminata Vietnam*; BLRaV, *Birch leaf roll-associated virus*; BVF, *Blackberry Virus F*; BsCVBV, *Bougainvillea spectabilis chlorotic vein-banding virus*; CBSLV, *Cacao bacilliform Sri Lanka virus*; CaMMV, *Cacao mild mosaic virus*; CSSCDV, *Cacao swollen shoot CD virus*; CSSCEV, *Cacao swollen shoot CE virus*; CSSGMV, *Cacao swollen shoot Ghana M virus*; CSSGNV, *Cacao swollen shoot Ghana N virus*; CSSGQV, *Cacao swollen shoot Ghana Q virus*; CSSTAV, *Cacao swollen shoot Togo A virus*; CSSTBV, *Cacao swollen shoot Togo B virus*; CYVBV, *Cacao yellow vein banding virus*; CaYMV-A, *Canna yellow mottle virus isolate*; CaYMV-Ci, *Canna yellow mottle-associated virus*; CYMV, *Citrus yellow mosaic virus*; CoVCV, *Codonopsis vein clearing virus*; CoYMV, *Commelina yellow mottle virus*; CyNLV, *Cycad leaf necrosis virus;* DBALV, *Dioscorea bacilliform AL virus*; DBALV2, *Dioscorea bacilliform AL virus 2*; DBESV, *Dioscorea bacilliform ES virus*; DBRTV1, *Dioscorea bacilliform RT virus*; DBRTV2, *Dioscorea bacilliform RT virus 2*; DBSNV, *Dioscorea bacilliform SN virus*; DBTRV, *Dioscorea bacilliform TR virus*; DrMV, *Dracaena mottle virus*; FBV-1, *Fig badnavirus-1*; GVBV, *Gooseberry vein banding virus*; GRLDaV, *Grapevine roditis leaf discoloration-associated virus*; GVCV, *Grapevine vein-clearing virus*; HiBV, *Hibiscus bacilliform virus*; JuMaV, *Jujube mosaic-associated virus*; KTSV, *Kalanchoe top-spotting virus*; MBV-1, *Mulberry badnavirus-1*; PYMAV, *Pagoda yellow mosaic associated virus*; PVBV, *Pelargonium vein banding virus*; PBCOV, *Pineapple bacilliform comosus virus*; PYMoV, *Piper yellow mottle virus*; RYNV, *Rubus yellow net virus*; SCBBBV, *Sugarcane bacilliform BB virus*; SCBBOV, *Sugarcane bacilliform BO virus*; SCBBRV, *Sugarcane bacilliform BRU*; SCBGAV, *Sugarcane bacilliform Guadeloupe A*; SCBGDV, *Sugarcane bacilliform Guadeloupe D virus*; SCBIMV, *Sugarcane bacilliform IM virus*; SCBMOV, *Sugarcane bacilliform Mor virus*; SPPV, *Sweet potato pakakuy virus*; SPBVb, *Sweetpotato badnvirus b*; TaBCHV, *Taro bacilliform CH virus*; TaBV, *Taro bacilliform virus*; WBV1, *Wisteria badnavirus 1*; YNMoV, *Yacon necrotic mottle virus isolate YV1*. *Rice tungro bacilliform virus* (RTBV) was used as an outgroup.

with *Citrus yellow mosaic virus* (CYMV) and *Hibiscus bacilliform virus* (HiBV) (Fig. 5). Regarding the phylogenetic distribution of badnaviruses and the class of plants they infect, the cacao-infecting badnaviruses belonged to the cluster of viruses characterized by infecting plants of the Magnoliopsida class. However, CAMMV, CBSLV and CYVBV were clustered with viruses that affect plants of both the Liliopsida and Magnoliopsida classes (Fig. 5).

3.7. Evaluation of temporal signals and molecular clocks in cacao-infecting badnavirus phylogeny

To eliminate possible problems of confounding temporal and genetic structures, before performing the temporal analysis, the Mantel confounding test [52] was performed. The absence of any confounding factors ($p = 0.096$) is favourable evidence of the temporal signal in the dataset (Supplementary materials 11 and 12). The results for the root-to-tip regression showed that the dataset used in this study had a positive temporal signal, with a correlation coefficient of 0.4279, R2 of 0.1831, root-to-tip p value of 3.56E-05 and evolutionary rate of 2.7174E-02 (Supplementary material 13). The real presence of a temporal signal in the data was corroborated by the permutation analysis, in which a significant p value for rejecting the null hypothesis (p value = 0.999) was obtained. This p value is the proportion of replicates with a statistical test greater than or equal to the true value. The null hypothesis corresponds to a negligible amount of evolution taking place between the sampling dates, so that the correlation observed can be attributed to stochastic variation in molecular branch length estimates and to rooting the tree to maximize clocklikeness (Supplementary material 11). Based on the positive evidence of a strong temporal signal in the available data, a molecular clock of cacao-infecting badnaviruses was built (Fig. 6). The first common ancestor to all cacao-infecting badnaviruses was located in 1775.6. From this point, the most recent shared ancestor between CBSLV and CYVBV emerged in 1807.3. Prior to this event, a shared ancestor in 1786.9 emerged between CaMMV and the African cacao-infecting badnaviruses. The separation between CaMMV and the African cacao-infecting badnaviruses took place in 1808.2, which corresponded to the date of the African species diversification events. The first direct descendant of the common ancestor dated from 1808.2 and was the CSSGQV species (1942.3; Fig. 6). According to the molecular clock, in 1876.6, a new common ancestor was identified and gave rise to the CSSGNV and CSSGMV species (whose closest ancestor appeared in 1922.4) and a new ancestor (1895.2, Fig. 6). From this last ancestor (1895.2), the tree diverged and originated the CSSCEV species (most recent ancestor 1948.8) and a new ancestor (1924.1, Fig. 6). From 1924.1, ancestors originated in 1940.8, which represented the shared ancestor of CSSTAV (1970) and CSSCDV (1992.4) and the CSSTBV species (1957). Thus, the ancestor of the cacao-infecting badnaviruses as well as the ancestor of CaMMV arose at the end of the 18th century. The ancestors of CYVBV, CBSLV, CSSGQV, CSSGMV and CSSCEV were located in the 19th century, while the ancestors of CSSTAV, CSSCDV and CSSTBV were located in the 20th century (Fig. 6).

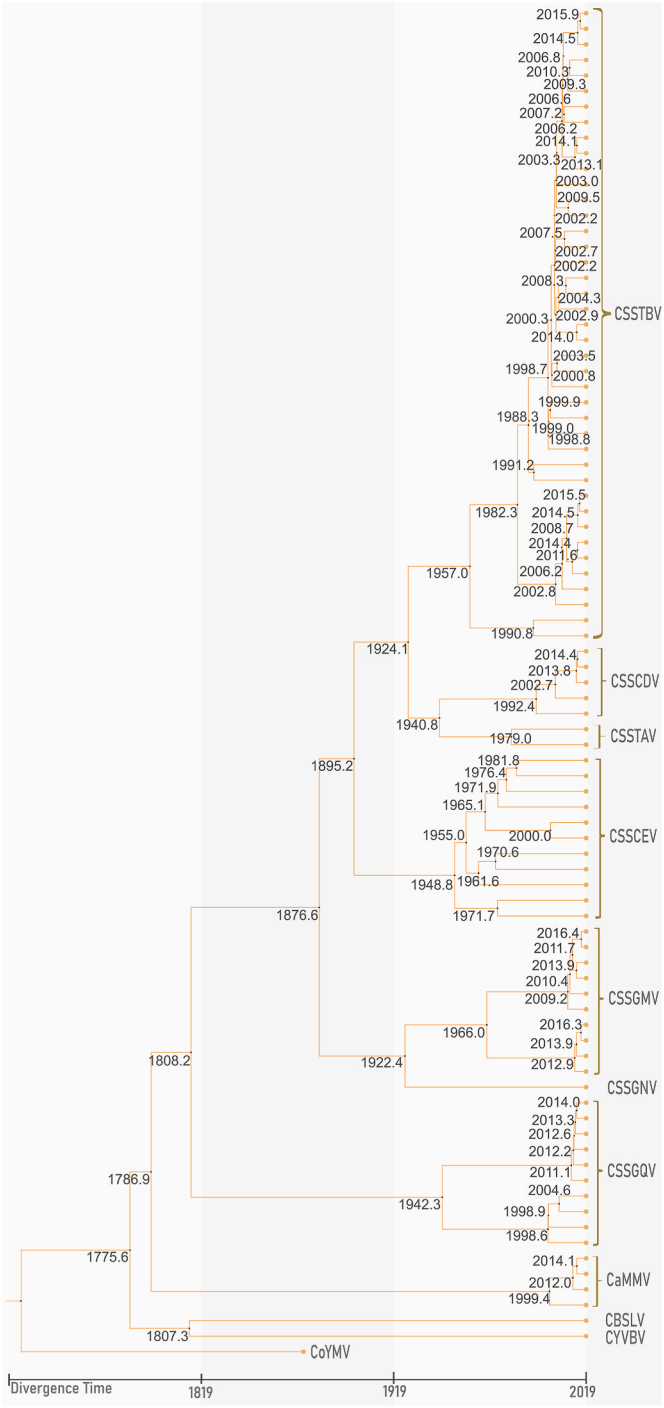


Fig. 6. Timetree analysis based on complete genomes phylogeny of cacao-infecting badnaviruses. The phylogeny was reconstructed using Maximum Likelihood method. Model Substitution GTR + F + I + G4. The timetree was inferred by applying the RelTime method, it was computed using 86 calibration dates (Table 1). Branch lengths are expressed in divergence times. Sampling times range from 1990 to 2019. CBSLV, *Cacao bacilliform Sri Lanka virus*; CaMMV, *Cacao mild mosaic virus*; CYVBV, *Cacao yellow vein banding virus*; CSSCDV, *Cacao swollen shoot CD virus*; CSSCEV, *Cacao swollen shoot CE virus*; CSSGMV, *Cacao swollen shoot Ghana M virus*; CSSGNV, *Cacao swollen shoot Ghana N virus*; CSSGQV, *Cacao swollen shoot Ghana Q virus*; CSSTAV, *Cacao swollen shoot Togo A virus*; and CSSTBV, *Cacao swollen shoot Togo B virus*. *Commelina Yellow Mottle Virus* (CoYMV) was used as an outgroup.

4. Discussion

4.1. Genetic diversity of cacao-infecting badnaviruses and its relationship with geographic region and symptoms variability

Historically, studies of the genetic diversity of African cacao-infecting badnaviruses suggested the existence of a possible relationship between Geographic region and viral genomes and the lack of correlation between Symptoms and genetic diversity [14,15,20,22,23]; however, these hypotheses were never rigorously tested. Thus, the analyses presented here demonstrated with statistical significance the existence of a strong phylogenetic signal both for the Geographical region and for the Symptoms described for each of the available genomes ($\lambda = 1$, Fig. 1).

According to the data collected here (Table 1, Fig. 1), Ghana is the country with the highest diversity of CSSD associated species (six species, Supplementary material 10) and the highest diversity of symptoms, which may be due to the diversity of species existing in that country. However, the largest number of genetic sequences analyzed came from Ivory Coast. In relation to the species of cacao-infecting badnavirus, those that presented the greatest global distribution were CSSTBV and CaMMV, thus evidencing the existence of species with greater geographical expansion than others, however, this may be due not only to human action, biological factors such as differences in the transmission capacity of each species, population and vector specificity among others [8,57,58], but also to the great imbalance that exists in the number of sequences available per country, so it is possible that the real diversity of countries where the number of sequenced genomes is really low is being underestimated, so it is considered advisable to carry out massive sequencing projects for the other species of cacao-infecting badnavirus, in order to reduce possible biases due to the low availability of sequenced genomes.

If it is considered that viral genomes may contain relevant information that influences virulence and pathogenicity processes, it is possible that there is a correlation between genetic diversity and these characteristics [32,59]. However, one of the most common failures when evaluating such correlations lies in the choice of the genetic sequence to correlate, repeatedly resulting in spurious correlations [60]. In the case of CSSD associated viruses, it has been shown that the use of complete genomes for correlation analysis generated considerably more robust results [18] than the use of partial sequences, such as the first part of ORF3, which led to good correlations only with the Geographic region and a total absence of correlation with the Symptoms [15]. Our results indicated that the phylogenetic arrangement of cacao-infecting badnaviruses species was correlated with the symptoms they trigger, i.e., within the group of CSSD associated species, it was possible to recognize and discriminate species that triggered more or less severe symptoms, and such differences could have a genetic basis (Fig. 1). This result corroborated previous data showing that although CSSV symptoms are variable from one strain to another, they are hereditary and maintained over time because plants generated from healthy cacao seeds infected with inoculum from diseased plants showed the same symptoms as the inoculum source plants [23].

Our data suggested three possible patterns linking genetic identity, Geographic region and Symptoms, for CSSD associated species: (i) high genetic identity accompanied by narrow geographic distribution and uniform symptoms; (ii) high genetic identity accompanied by wide geographic distribution and diverse symptoms; and (iii) low genetic identity, wide geographic distribution and uniform symptoms (Fig. 1, Supplementary material 8).

Pattern (i) coincides with the classical idea of the species, where a high genetic identity is associated with restricted geographical areas and uniform symptoms, and it even recalls processes of endemism, such as in the case of the CSSGMV species [21]. A similar phenomenon was observed in *Human immunodeficiency virus* (HIV), for which the lower geographic expansion of certain viral groups was linked to their low genetic variability [61]. On the other hand, the CSSGQV species also framed in this pattern, probably due to its symptomatic peculiarity and phylogenetic position (Fig. 1). It behaves similarly to a defective viral genome or mutant genome, whose the completeness functionality generally depends on the complementarity of the genomes of the group [62]. Due to their nature, these viral particles are characterized by a low geographical distribution and participation in coinfection processes [63]. Both characteristics were observed in the CSSGQV species, and recent research shows that this type of behavior occurs not only in RNA viruses but also in DNA viruses [64].

Pattern (ii) was observed for CSSTBV and CSSTAV species, and it was previously reported by Refs. [15,65]. In both cases, the detection of this pattern has been considered an atypical and confusing phenomenon, which has led to the base argument for rejecting the existence of a relationship between symptomatology and viral genetic diversity. However, it is necessary to highlight the complexity and multiplicity of factors involved in viral infections because this kind of pattern was observed in species with a wide geographic distribution, which could greatly modulate the observed symptoms [66,67]. In addition, the presence of more or less virulent variants within the same viral species is a widely accepted phenomenon; in the case of *Maize Streak Virus* (MSV), various symptoms from mild to severe were reported associated with variants that shared a genetic identity of 95% [68].

Finally, pattern (iii) is of special interest because it is mainly associated with the CSSCEV species, the most lethal species associated to CSSD described thus far. The CSSCEV species is associated with rapid death syndrome in cacao trees, a worrying phenomenon in which the lifespan of the plant from the onset of infection is shortened from 3 to 5 years to only 1 year [65]. Here, we suggest that the high degree of genetic diversity observed in CSSCEV species could influence the severity of symptoms due to the behavior of the viral populations as a quasispecies [69]. Thus, the cloud of mutants that comprises a quasispecies sometimes seems to behave as a whole, where the parts interact in processes of complementarity and cooperation [70]. Indeed, compared with the selection of individual adaptive mutations, the greater quasispecies diversity per se is responsible for increasing pathogenicity and virulence and generating more severe and even lethal symptoms. It is believed that such a phenomenon is due precisely to the effect of cooperative and complementarity interactions experienced by the different variants within the quasispecies [62,71,72]. Moreover, in animal RNA viruses, increased lethality has been shown in response to an increase in the degree of genetic variability of the quasispecies [73]. In addition, a wide geographic distribution, as observed for CSSCEV species, has been associated with higher virulence levels in several pathogens, including MSV [68,74].

4.2. Evolution of the geographic region and symptoms of CSSD associated species

In reconstructing a possible scenario of evolution of the geographical distribution of cacao-infecting badnaviruses on the African continent, it was encouraging to find that the history written in base pairs in the viral genomes coincides quite well with the chronicle described for the development of the epidemic. Therefore, the model with the best fit (SYM_EB; Table 2) suggests that the beginning of the spread of this virus occurred in Ghana or Nigeria, and it later moved to the Ivory Coast and finally to Togo (Fig. 2A), which coincides with the historical records of virus dispersal [7,12,13,75]. In addition to the above, the EB transformation describes adaptive radiation processes, characterized by a rapid initial diversification that slows down over time [42]. The initial explosion has been associated in many cases with the use and colonization of new areas because the slowdown may be due to competition processes and limitations in the colonization of new areas, among many other factors [76]. In the particular case of the geographical distribution of CSSD associated species in Africa, the deceleration process ($a = -7784$), which can be interpreted as a decrease in the timescale of the geographical spread of the viruses on the African continent, was consistent with the various initiatives adopted in different countries to stop the geographical expansion of the viruses (e.g., “cutting-off” campaigns, adoption of phytosanitary controls and greater regulation in the transit of plants) [28,77].

The reconstruction of ancestral states allowed us to corroborate the hypothesis of the possible initial emergence of the CSSCEV species between Ghana and the Ivory Coast (Fig. 3) [18]. Furthermore, the reconstruction presented here suggests that Togo and the Ivory Coast were possible centers of emergence of CSSTBV, while the emergence of CSSCDV may have exclusively occurred in the Ivory Coast, constituting a new contribution in relation to the biogeographic history of CSSD associated species (Fig. 3). For American cacao-infecting badnaviruses, new contributions have been made, with Brazil identified as a possible

ancestral region of CaMMV (Supplementary material 6), which is totally new and of great interest in the reconstruction of the evolutionary history of cacao-infecting badnaviruses.

In relation to the study of the Symptom evolution, it is important to mention that historically, phytopathological studies have primarily focused on the evolution of pathogenicity, which determines the ability of a pathogen to cause a disease in a particular host, and paid less attention to the evolution of virulence, which is understood as the degree of damage caused to the host [78]. In virus-plant interactions, it is assumed that symptom expression is a proxy for virulence, which has indeed been shown for several virus-plant interactions [79]. For this reason, in the present study, we took the symptoms described for each CSSD associated viral genome as an approximation of the virulence of each genome. The ordered evolution model (Meristic_EB, which presented the best fit; Table 2) suggested that CSSD associated viruses would have evolved from less to more severe symptoms (Fig. 2B), which is compatible with the historical records of the development of the epidemic in Nigeria [21]. The ancestry analyses (Fig. 4) support this approach, despite presenting difficulties in accurately predicting the symptoms of the root ancestors, which may be due to the recent evolutionary history of the viruses, as is also reported for the ancestral reconstruction of epidemic locations of the virus dengue (DENV2), where the power and specificity of the prediction of ancestral states increased with the progression through the tree and moved away from the root node [49].

Consequently, it appears that the symptoms (although variable) triggered by the different CSSD associated isolates may have a genetic basis (Fig. 1), which contradicts the ideas established in previous research works [14,15,20,22]. Stem swelling is a symptom that deserves to be analyzed in detail because it has historically been considered the distinguishing symptom of the disease and has been used to differentiate and delimit American from African cacao-infecting badnaviruses. However, according to the Meristic_EB model and ancestral reconstruction (Supplementary material 5 and 7, respectively), this symptom appeared and disappeared in the evolution of the symptoms considered here, which is actually in agreement with the widely documented observations in several outbreaks in Ghana, Nigeria, Ivory Coast and Sri Lanka [8,20,21,23,40,80–84]. All of these results suggested that the expression of stem swelling is highly variable and that its absence did not imply lower severity of the disease; indeed, a significant decrease in canopy and production was recorded in cacao plants that did not show stem or root swelling despite being infected by these viruses [82,85–89]. It should be emphasized that the evolutionary model proposed here is relevant in that it provides a broader and more integrative view of the history of this disease, however, as has been expressed by different researchers, predicting the evolution of virulence is extremely complex and requires numerous studies, because each pathogen, the host and the circumstances of both are unique and defining in the evolutionary process [90].

4.3. New insights into the origin of cacao-infecting badnavirus species

For decades, the absence of detailed information on cacao-infecting badnavirus outside the African continent, together with the lack of studies with new approaches, favored the controversy regarding the origin of CSSV and allowed the establishment of one of the possible hypotheses as true, however, and given the importance of the subject, we consider it appropriate to perform here a detailed analysis of each of the possible hypotheses raised so far.

Hypothesis 1. Africa as the center of origin of the virus

Although the first reported date of CSSV in Ghana is 1922, by that time the disease was already so widespread and so severe that several sources place the actual beginning of the epidemic several years earlier, probably in 1907 [91]. However, it was not until 1940 that it was demonstrated that the causal agent of the disease was a virus, promptly suggesting that the origin of the pathogen must be in Africa, because the swollen shoot disease of cacao was first described in this continent and until then was completely unknown in America, the original source of cacao [92]. In view of this hypothesis, it was proposed that plants native to the African continent would be the original source of the viral particles, so Chevalier (1946) [93] suggested the evaluation of the taxonomically closest relatives of *T. cacao* and natives of Africa as possible natural reservoirs of the infection, highlighting three candidate species: *Leptonychia urophylla* (=*L. pubescens* Keay), *Buettneria africana* Mast (= *Byttneria catalpifolia* Jacq.) and *Scaphopetalum amoenum* A. Chev. However, it would not be until the 1950s when the first investigations would begin, highlighting the studies of Posnette et al. (1950), Todd (1951) and Tinsley & Wharton (1958) [83,94,95], works that would lay the foundations of the idea of an African origin of the CSSV. The main findings of these investigations could be summarized as follows:

(i) Posnette et al. (1950) [83] identified 7 plants susceptible to CSSV, highlighting among all of them three species *Cola chlamydantha* K. Schum, belonging to the family Sterculiaceae, *Ceiba pentandra* Gaertn. and *Adansonia digitata* Linn, belonging to the family Bombacaceae. The above species showed the greatest susceptibility to infection by infected mealybugs, however, and as the authors mentioned, these species are more difficult to infect than cacao and the rate of transmission of the virus from them to cacao through mealybugs is considerably low, in many cases nonexistent, noting that mealybugs do not seem to become infectious easily when they feed on infected native plants, so they do not effectively transmit the virus to cacao. In addition, the authors suggest that *A. digitata*, despite being the most susceptible plant and with the highest rate of virus transmission to cacao, is an unlikely original source of the virus, as it is native to dry areas unsuitable for cacao cultivation, therefore, they propose *C. chlamydantha* and *C. pentandra* as possible sources of origin of the virus. Among the two, *C. pentandra* was the only plant located in Nigeria susceptible to the virus, however, the finding of *C. chlamydantha* plants near Wiawso in the Western Province of Ghana with symptoms of infection and capable of transmitting the virus to cacao beans, caused this plant to be chosen as the most plausible option of original reservoir of the CSSV in Africa.

(ii) The year following the publication of the work of Posnette et al. (1950) [83], *C. chlamydantha* was already referred to as the original source and natural reservoir of CSSV, so the observations of Todd (1951) [94], by expanding the record of infected specimens in regions both near and further away from cacao plantations in the aforementioned Wiawso region as well as in the districts of Dunka, Enchi, Prestea, Axim, Sekondi and Cape Coast, all belonging to the Western Province of Ghana, constituted only a confirmation of the approach proposed by Posnette et al. (1950) [83], from which it was suggested that CSSV viral infection in *C. chlamydantha* constituted an endemism.

(iii) At the end of the decade, Tinsley & Wharton (1958) [95] expanded the evaluation of possible hosts indigenous to Africa, concentrating their research exclusively on plants of the families Tiliaceae, Sterculiaceae and Bombacaceae, evaluating the susceptibility of twenty seven plant species to eight different strains of the virus, which at that time constituted all known strains of CSSV. Similar to studies (i) and (ii) Tinsley & Wharton (1958) [95] identified that plants belonging to the Bombacaceae family present higher susceptibility to the virus than plants of the Sterculiaceae family, in fact, while *C. pentandra* and *A. digitata* are susceptible and were experimentally infected by seven and eight CSSV variants, respectively; *C. chlamydantha* was susceptible to only three of the eight virus variants evaluated. Likewise, the authors reported that the plants proposed by Chevalier (1946) [93] were not infected by any of the viral variants evaluated, so they could not be a reservoir and original source of the virus.

However, the proposal that *C. chlamydantha* was the original source of the virus was strengthened until the basis of an African origin of CSSV was established. This idea was maintained over time and since the advent of new sequencing technologies, it was even raised the possibility that each group of genetic diversity of the virus could have arisen independently and spontaneously from different wild hosts in each of the countries of West Africa [20]. This hypothesis, however, needs to be tested because to date, and this is extremely relevant to consider, there is no genetic evidence of CSSV diversity in any of the possible indigenous hosts mentioned above. In fact, there is no available complete genome of a cacao-infecting badnavirus from these indigenous hosts; only fragments of viral genomes isolated from *C. pentandra* are in the data banks [22].

Hypothesis 2. America as the center of origin of the virus

In the 1970s, new research would lead to a re-evaluation of the geographical origin of CSSV. Indeed, by thoroughly evaluating the available data on host range, symptoms and transmission by different vectors of the eight varieties of CSSV described so far, Bald & Tinsley (1970) [96] found important correlations between these variables, which allowed them to conclude that despite what had been established until then, *C. chlamydantha* was not the main and original host of CSSV, since this plant was actually infected from cacao. Some of the main evidences that support such a conclusion are the geographical distribution of *C. chlamydantha*, the CSSV variants capable of infecting this plant and the specificity of certain vectors, since when evaluating the origin of the *C. chlamydantha* specimens reported by Posnette et al. (1950) [83] and Todd (1951) [94], which in principle are the evidence supporting the hypothesis of an African origin of the virus, it was determined that they came from western Ghana and could be infected with a small number of CSSV variants, which also generated mostly severe symptoms [96]. In contrast, and as observed in Fig. 2B and 4, the historical records indicate that the disease symptoms associated with the first outbreak were mild and would have occurred at the opposite end of Ghana, in the eastern region, where, in fact, cacao would have been introduced and the first plantations established; which is why one of the initial outbreaks, perhaps the first and largest, is the famous Mampong outbreak, an outbreak at least 30 miles in diameter, which even included some of the first cacao trees introduced into West Africa [97].

Furthermore, the CSSV variants associated with the Mampong outbreak are totally incapable of infecting *C. chlamydantha* plants, and while the variants capable of triggering infection in this plant generate severe symptoms when evaluated in cacao seeds, the original Mampong variants generate different symptoms, because even when they initially appear severe, the symptoms diminish rapidly until the cacao plants become totally asymptomatic, as a result of the above, Bald & Tinsley (1970) [96] concluded that the absence of *C. chlamydantha* at the site of origin of the disease, the impossibility of the oldest strains of CSSV to infect this plant, as well as the specificity of the vector that transmits them and the evolution of the symptoms, suggest that it is incorrect to associate the origin of CSSV to *C. chlamydantha*; on the contrary, this plant possibly acquired the disease from diseased cacao that probably moved from eastern Ghana (where cultivation began) to the west, which would also explain the evolution of the virus by increasing the severity of symptoms and the number of vectors.

Similarly, Tinsley (1971) [98] highlights the absence of a consistent association between infected indigenous hosts and the incidence of the disease, suggesting that indigenous plants, such as *C. chlamydantha*, would be of local importance, but could never be considered as original hosts of the virus, since the evidence supporting the widespread view that the swollen shoot virus is of West African origin is equivocal, as quoted by the author: "Nevertheless, the evidence available to incriminate the wild vegetation of the West African forest as the original source of the swollen shoot virus is not convincing. This raises the possibility of another geographical origin for the swollen shoot virus […] therefore, it would be realistic to examine the possibility that this cacao virus has a New World origin."

In fact, the molecular clock carried out in the present study (Fig. 6) provided evidence for reevaluating established ideas and strengthened the hypothesis that the emergence of cacao-infecting badnaviruses could have occurred in America because the first diversification event that groups the current species was dated to the 18th century, which was long before the introduction of cacao in Africa (Fig. 7). Indeed, recent analyses demonstrated the insertion of badnavirus sequences in most of the cacao genotypes established by Motamayor et al. (2008) [99], and this insertion was considered quite ancient and associated with the time of diversification of cacao [100]. Consequently, if the center of origin and diversification of cacao is located in South America in the upper Amazon basin [1,2], the presence of badnaviruses capable of infecting cacao in that region is also expected. In addition, new evidence could support this hypothesis:

(i) The existence of a common ancestor between CYVBV and CBSLV dated to 1808 (Fig. 6), which coincided with historical data stating that cacao was introduced to Sri Lanka from America and not from Africa. In 2014, Frasch [101] suggested that the first cacao plants were introduced to Ceylon (former name of Sri Lanka) through Dutch settlers before 1822. Other sources placed the introduction of cacao in Sri Lanka in 1834, coming from Venezuela, which is also the place of origin of cacao brought to Trinidad in 1525 and one of the major centers of cacao cultivation and dispersion in colonial times [5,102,103]. In any case, both hypotheses support the existence of a common and ancestral geographic point located in America, which could link the emergence of both plants and viruses (Fig. 7) and explain the notorious absence of kinship between CBSLV and the African cacao-infecting badnaviruses, as evidenced in all the analyses performed here (Figs. 1 and 5). In line with the aforementioned, Ullah and Dunwell [104] while studying endogenous badnavirus sequences in the genome of *T. cacao*, confirms the existence of a potential phylogenetic link between CYVBV and CBSLV.

(ii) The American CaMMV species shared a common ancestor with the CSSD African associated species, indicating that CaMMV was genetically closer to African CSSV species than to CYVBV, despite the localization of CYVBV in Trinidad and Tobago (Figs. 1 and 6). According to the reconstruction of ancestral states, Brazil (South of Bahia State) would be the possible ancestral territory of the current CaMMV sequences (Supplementary material 6) [105]. Cacao cultivation in Bahia State began in 1746, when a French colonizer living in Pará (northern state of Brazil) sent some seeds of the "Forastero" variety (Amelonado group) to a landowner located in southern Bahia (Fig. 7) [106]. Young (1994) [103] located the introduction of cacao in Bahia State four years later, in 1750. Both dates were prior to the emergence of the shared ancestry between CaMMV and African badnaviruses (1786.9, Fig. 6), which is of utmost interest since southern Bahia State was the origin of cacao plants (Amelonado type) introduced in Africa in 1822 (Fig. 7) [6]. Therefore, it would seem that this common ancestry between the American and African species could have originated in America. The recent detection of CaMMV and CYVBV sequences in cacao accessions found in South America, the Caribbean and Southeast Asia supported our observations and hypotheses (i) and (ii), however, the authors of this research suggest the CaMMV as a possible spreader of the disease in Java and Southeast Asia, while here, we showed that CYVBV is more closely related to the only fully sequenced cacao-infecting badnavirus in Asia, CBSLV (Fig. 1), in the same manner, the permanence of CYVBV and CaMMV in asymptomatic cacao plants could have favored their dispersal in various countries of America and worldwide [107].

(iii) The most recent shared ancestor between CSSGQV (first species of cacao-infecting badnavirus to appear in Africa; Fig. 3) and the

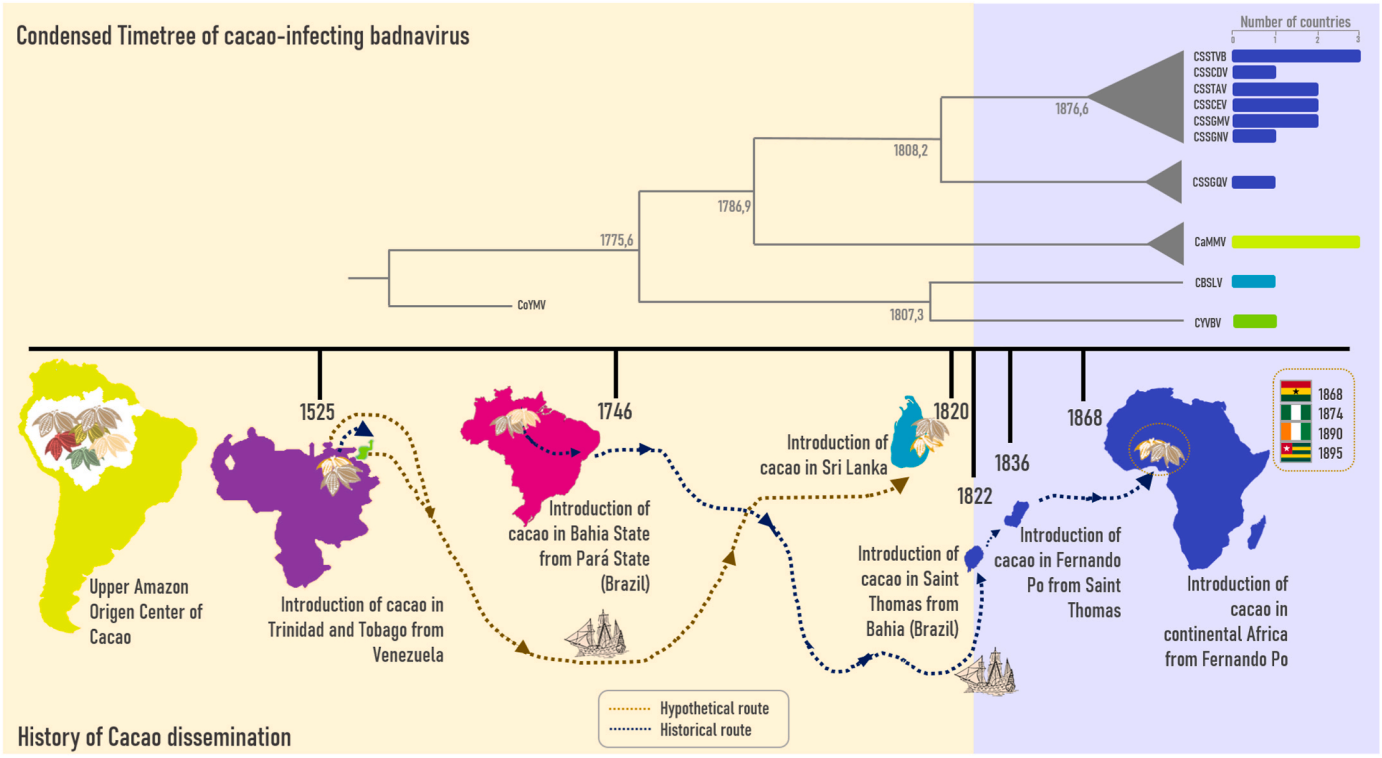


Fig. 7. General scheme of cacao-infecting badnaviruses evolution through time associated with history of cacao dissemination.

other African species was dated to 1808 (Fig. 6), which was before the introduction of cacao to the African islands of Saint Thomas and Prince (1820–1822) and Fernando Po (1836–1854) then, it is suggested that this ancestor could also have originated in America (Fig. 7) [6,108–113].

(iv) The major diversification event of African species was dated to 1876.6 (Fig. 6), which coincided with the introduction of cacao to mainland Africa from Fernando Po Island, first in Ghana 1868–1879 and later in Nigeria in 1874 (Fig. 7) [6,103]. The two species derived from this initial ancestor (possibly the first African) were CSSGNV from Ghana (more basal) and CSSGMV from Nigeria (Fig. 1), thus coinciding with the predictions made by our evolutionary model (Fig. 2A). Regarding the origin of African cacao, both the historical records and its homogeneity - 95 % of all cacao grown on the continent up to the 1940s was Amelonado type - support the hypothesis of one or at most two introductions from the same place [6,114]. Therefore, it is very likely that both Gold Coast and Nigerian cacao came from a common source: Fernando Po; since the cacao introduced to this island came entirely from Brazil, it is considered that the origin of African cacao is in Brazil [6]. In agreement with the above, the presence of CYVBV and CaMMV in several asymptomatic cacao plants may contribute to explain the entry of the virus into Africa from plants coming from America, in fact, the finding in 2021 of plants infected with CaMMV in plantations in Brazil, being that this virus was even considered eradicated from Trinidad and Tobago when it was detected on the island in 2017, it is an example of how the virus can remain undetected in American ecosystems and therefore give rise to the misconception that the disease did not exist in America before its description in Africa [16,100,105].

(v) Considering the phylogenetic relationships shown in Figs. 1 and 6, it seems unlikely that each of the CSSD associated species originated in a different indigenous host, which would also imply that in a short time in two distant geographical regions such as Ghana and Nigeria (original foci of the disease in Africa, according to historical data and supported by the results presented in Fig. 2A and 3), two jump events from indigenous plants to cacao would have to have occurred completely independently and spontaneously and both viruses would have to be morphologically and genetically similar, this juncture would also apply to Sri Lanka (Asia) where the virus would also have to have arisen spontaneously from an indigenous host and share the same morphological, genetic and symptomatic characteristics of African viruses in a completely coincidental way, since there is no evidence of the introduction of cacao in Sri Lanka from Africa and if from America, therefore, the original source of the virus in each country would be different, resulting in an unparsimonious hypothesis [98,115].

The apparent absence of the virus in the America could be due to several factors. First, the high genetic diversity of cacao in its place of origin could have functioned as a buffer against viral infections, and there is ample evidence of how genetic diversity allows a greater degree of adaptation and host response to infection [116–118]. Second, the growing conditions and management in America were radically different from those in Africa, greater genetic variability and lower susceptible host density may have contributed to lower virus transmission rates [74,119,120]. In addition, full sun cultivation implemented in Africa is known to directly impact the severity of the induced symptoms by CSSD associated viruses, as plants exposed to full sun and infected by CSSD associated viruses exhibit stem swelling and other severe symptoms of the disease, including increased mortality. Fifty percent coverage, in contrast leads to disappearance of severe symptoms and predominance of mild symptoms similar to those described in America for CYVBV and CaMMV, where cacao cultivation was mostly in shade [16,121–123]. Third, it should be noted that there are no published cacao badnavirus studies in Colombia, Venezuela, Dominican Republic and Brazil, being that for these countries the presence of plants with viral symptoms was previously reported [122–124]. The most recent studies in this area are related to the identification of polerovirus

in *T. cacao* germplasm from Costa Rica, although it is an RNA virus [124].

It is equally necessary to emphasize the crucial role that vectors may have played in the development of the disease in both Africa and America, considering significant differences in the composition of mealybug species in each region [57,125–127]. In Ghana, for example, the dominant species associated with *T. cacao* is *Formicococcus njalensis*, which is up to a hundred times more abundant than the second most common species (*Planococcus citri*). Conversely, in America, *Ferrisia virgata*, *Pseudococcus comstocki*, *Pseudococcus brevipes*, *Pseudococcus citri*, *Pseudococcus jackbeardsleyi*, *Pseudococcus viburni* and *Maconellicoccus hirsutus* are the most abundant [57,128–130]. The abundance and richness of mealybugs per se, perhaps may not be a particularly crucial differentiator in the adaptive radiation of cacao-infecting badnaviruses in Africa, however, there are certain key details that when considered, may indeed contribute to explaining the phenomenon experienced in this region.

In fact, *F. njalensis* the dominant species in Africa, is considerably more efficient in the acquisition and transmission of viral particles than the other mealybug species evaluated in both regions. This behavior is partially attributed to the possession of a larger stylet, enabling *F. njalensis* to access both the stem phloem and leaf mesophyll. In contrast, the smaller stylet of the American species *F. virgata*, for example, can hardly reach the phloem and remains restricted to the leaf mesophyll. This difference could be considered almost transcendental, as the highest concentration of viral particles in *T. cacao* plants is precisely in the phloem, not the mesophyll. Therefore, the African species *F. njalensis* is indeed capable of acquiring a greater number of viral particles [131–133]. Additionally, it has been reported that, unlike other mealybug species that exhibit a semi-persistent transmission mechanism for cacao-infecting badnaviruses, *F. njalensis* appears to possess a persistent transmission mechanism, this characteristic could further enhance the transmission capacity of this vector [129,134].

While there appears to be a greater capacity for the acquisition and transmission of viral particles by some of the dominant mealybug species in African countries, the cultivation conditions in this region may have also contributed to the expansion of vectors, thus contributing to the severity of the epidemic. In fact, there is evidence that full-sun cultivation of *T. cacao* can influence both the feeding habits of mealybugs and the population dynamics of ants associated with them. This cultivation practice also reduces protection against the wind, further facilitating vector dispersion [135,136].

Therefore, differences at the vector level may contribute to explaining the observed disparities in the evolution of cacao-infecting badnaviruses between one continent and another, in addition to this, other important factor that contributed to explaining the rapid diversification of cacao-infecting badnaviruses was based on their nature: The *Badnavirus* genus is characterized as the most recently evolved and the genus with the highest nucleotide substitution rates per site within the family *Caulimoviridae* [137]. Indeed, according to the analyses carried out here (see §3.7), the substitution rate for CSSD associated viruses is $2.7174x10^{-2}$ (substitutions/site/year). This substitution rate is similar to that reported for *Banana streak MY* virus, a widely studied badnavirus, which presents a whole genome substitution rate of $1.48x10^{-2}$ (substitutions/site/year) [138]. Therefore, we consider that the value obtained for CSSD associated viruses is comparable to the ranges described for other viruses of the same genus, being these in turn comparable to the values proposed for RNA viruses (10^{-2}-10^{-5}) [90,139,140]. In fact, there is evidence of dsDNA viruses with substitution rates of up to 10^{-5}, as this type of virus is considered the most conservative and with lower substitution rates, it is expected that plant pararetroviruses of similar behavior to animal retroviruses present higher substitution rates than dsDNA viruses [141,142].

These results, which were obtained with exhaustive and robust temporality analyses (see §3.7) and the robust constructed phylogeny (Supplementary material 9), allowed us to hypothesize that cacao-infecting badnaviruses originated in America, which would explain both the presence of viral inserts in the different cacao genotypes [100, 104] as well as the phylogenetic (Figs. 1 and 6) and historical (Fig. 7) relationships between CYVBV-CBSLV and CaMMV-African CSSD associated species. Subsequently, the introduction of cacao in Africa along with the cacao-infecting badnaviruses, either through asymptomatic plants or with mild symptoms [107], would have constituted a true bottleneck, where a small population of badnavirus would develop the role of founders in the future.

5. Conclusions

Variability is possibly one of the main characteristics of cacao-infecting badnaviruses. This group of pararetroviruses has a high mutation rate, which could influence the different degrees of genetic diversity observed in the species that make up the group. Likewise, both the Geographical region and the Symptoms associated with each species showed a clear phylogenetic signal. Regarding the evolution of the aforementioned characteristics, the fit analyses suggested that the Geographical region of these viruses on the African continent evolved according to a symmetrical model (SYM_EB), with Ghana proposed as the center of virus dispersal, followed by Nigeria, Ivory Coast and Togo, which coincided with historical records. The evolutionary model that best fit the Symptoms described for each species was the ordered model (Meristic_EB), suggesting a general evolution of symptoms from less to more severe, starting with atypical symptoms (capable exclusively of coinfection) and proceeding to severe symptoms. Finally, the molecular clock provided new evidence in favour of an American origin of the cacao-infecting badnaviruses and provided a broader vision of the speciation process of these viruses in the African continent. All of these factors contributed to clarifying the evolutionary history of the group, thereby providing opportunities for future research.

CRediT authorship contribution statement

Patricia Lorena Gonzalez Grande: Conceptualization, Formal analysis, Investigation, Visualization, Writing – original draft, Writing – review & editing. **Fabienne Micheli:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

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

Data availability

All the data are available in the manuscript and in the supplementary material

Acknowledgments

PLGG was funded by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) through the Organization of American States and the Coimbra Group of Brazilian Universities scholarship programs. FM received a Productivity Grant from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). The authors thank the Centro de Computação Avançada e Multidisciplinar (CCAM, UESC) for providing computational infrastructure, Dr Darren Martin (University of Cape Town) and Dr Eric Roberto Guimarães Rocha Aguiar (UESC) for their comments and suggestions when this project was still an idea, Dr Christopher Griffin (Yale University) and Jean-François Dufayard (Cirad) for their help and comments on the evolutionary model analyses performed, Dr Emmanuelle Muller (Cirad) for providing information

about harvesting CSSV dates, and to Caio Suzart Argolo (UESC) for technical support.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.pmpp.2024.102242.

References

[1] J.E. Richardson, B.A. Whitlock, A.W. Meerow, S. Madriñán, The age of chocolate: a diversification history of Theobroma and Malvaceae, Front Ecol Evol 3 (120) (2015).

[2] S. Zarrillo, N. Gaikwad, C. Lanaud, T. Powis, C. Viot, I. Lesur, O. Fouet, X. Argout, E. Guichoux, F. Salin, R.L. Solorzano, O. Bouchez, H. Vignes, P. Severts, J. Hurtado, A. Yepez, L. Grivetti, M. Blake, F. Valdez, The use and domestication of *Theobroma cacao* during the mid-Holocene in the upper Amazon, Nat Ecol Evol 2 (12) (2018) 1879–1888.

[3] Faostat, Crops and Livestock Products Data, Food and Agriculture Organization of the United Nations, 2019.

[4] P.A. de Souza, L.F. Moreira, D.H.A. Sarmento, F.B. da Costa, Cacao—*Theobroma cacao*, in: S. Rodrigues, E. de Oliveira Silva, E.S. de Brito (Eds.), Exotic Fruits, Academic Press, 2018, pp. 69–76.

[5] G. Bondar, in: T.N. Bahia (Ed.), O Cultivo Do Cacau, 1956.

[6] F.N. Howes, The early introduction of cocoa to West Africa, Afr. Aff. 45 (180) (1946) 152–153.

[7] A. Posnette, Virus diseases of cacao in west Africa: i. cacao viruses 1A, 1B, 1C and 1D, Ann. Appl. Biol. 34 (3) (1947) 388–402.

[8] J. Thresh, T. Tinsley, The Viruses of Cacao, Weast African Cocoa Research Institute., 1959, pp. 5–32.

[9] H. Lot, E. Djiekpor, M. Jacquemond, Characterization of the genome of cacao swollen shoot virus, J. Gen. Virol. 72 (1991).

[10] A.I. Bhat, T. Hohn, R. Selvarajan, Badnaviruses: the current global scenario, Viruses 8 (6) (2016) 177.

[11] H. Alibert, G. Mangenot, A. Basset, Sur les caractères du " Swollen-Shoot " en Côte-d'Ivoire, J. Agric. Tradit. Bot. Appl. (283–284) (1946) 173–184.

[12] M. Partiot, Y.K. Amefia, E.K. Djiekpor, K.A. Bakar, Le "swollen shoot" du cacaoyer au Togo. inventaire préliminaire et première estimation des pertes causées par la maladie, Cafe Cacao The 22 (3) (1978) 217–228.

[13] J. Thresh, The control of cacao swollen shoot disease in Nigeria, Tropical Agriculture (Trinidad y Tobago) 36 (1) (1959) 35–44.

[14] E. Muller, S. Sackey, Molecular variability analysis of five new complete cacao swollen shoot virus genomic sequences, Arch. Virol. 150 (1) (2005) 53–66.

[15] F. Oro, E. Mississo, M. Okassa, C. Guilhaumon, C. Fenouillet, C. Cilas, E. Muller, Geographical differentiation of the molecular diversity of cacao swollen shoot virus in Togo, Arch. Virol. 157 (3) (2012) 509–514.

[16] N. Chingandu, M. Zia-Ur-Rehman, T.N. Sreenivasan, S. Surujdeo-Maharaj, P. Umaharan, O.A. Gutierrez, J.K. Brown, Molecular characterization of previously elusive badnaviruses associated with symptomatic cacao in the New World, Arch. Virol. 162 (5) (2017) 1363–1371.

[17] E. Muller, S. Ravel, C. Agret, F. Abrokwah, H. Dzahini-Obiatey, I. Galyuon, K. Kouakou, E.C. Jeyaseelan, J. Allainguillaume, A. Wetten, Next generation sequencing elucidates cacao badnavirus diversity and reveals the existence of more than ten viral species, Virus Res. 244 (2018) 235–251.

[18] R. Ramos-Sobrinho, N. Chingandu, O.A. Gutierrez, J.P. Marelli, J.K. Brown, A complex of badnavirus species infecting cacao reveals mixed infections, extensive genomic variability, and interspecific recombination, Viruses 12 (4) (2020).

[19] N. Chingandu, K. Kouakou, R. Aka, O.A. Gutierrez, J.K. Brown, Unexpected genome variability at multiple loci suggests cacao swollen shoot virus comprises multiple, divergent molecular variants, J Emerg Virol Dis 3 (1) (2017).

[20] K. Kouakou, B.I. Kébé, N. Kouassi, S. Aké, C. Cilas, E. Muller, Geographical distribution of cacao swollen shoot virus molecular variability in Côte d'Ivoire, Plant Dis. 96 (10) (2012) 1445–1450.

[21] N. Chingandu, L. Dongo, O.A. Gutierrez, J.K. Brown, The previously unidentified, divergent badnavirus species cacao red vein-banding virus is associated with cacao swollen shoot disease in Nigeria, Plant Dis. 103 (6) (2019) 1302–1308.

[22] F. Abrokwah, H. Dzahini-Obiatey, I. Galyuon, F. Osae-Awuku, E. Muller, Geographical distribution of cacao swollen shoot virus molecular variability in Ghana, Plant Dis. 100 (10) (2016) 2011–2017.

[23] J. Thresh, T. Tinsley, Virus Diseases of Cacao: the World Situation, Séptima Conferencia Interamericana de Cacao, Palmira-Colombia, 1958.

[24] G.A. Ameyaw, H.K. Dzahini-Obiatey, O. Domfeh, Perspectives on cocoa swollen shoot virus disease (CSSVD) management in Ghana, Crop Protect. 65 (2014) 64–70.

[25] B.S.H.B. Guiraud, M.G. Tahi, O. Fouet, C.I. Trebissou, D. Pokou, R. Rivallan, X. Argout, K.K. Koffi, B. Koné, B.I.A. Zoro, C. Lanaud, Assessment of genetic diversity and structure in cocoa trees (*Theobroma cacao* L.) in Côte d'Ivoire with reference to their susceptibility to Cocoa swollen shoot virus disease (CSSVD), Tree Genet. Genomes 14 (4) (2018) 52.

[26] E. Muller, Cacao swollen shoot virus, in: S.L. Rao, G.P. Khurana, S.M.P. Lenardon (Eds.), Characterization, Diagnosis and Management of Plant Viruses: Industrial Crops, 2008, pp. 403–409.

[27] S.Y. Agyeman-Boaten, A. Fumey, Effects of cocoa swollen shoot virus disease (CSSVD) on the welfare of cocoa farmers in Ghana: evidence from Chorichori community of the Sefwi Akontombra district, SN Business & Economics 1 (11) (2021) 149.

[28] C. Andres, A. Gattinger, H.K. Dzahini-Obiatey, W.J. Blaser, S.K. Offei, J. Six, Combatting cocoa swollen shoot virus disease: what do we know? Crop Protect. 98 (2017) 76–84.

[29] M.J. Jeger, L.A. Ollenu, L.S. Hagen, M. Jecquemond, Cocoa, in: T.G. Loebenstein G (Ed.), Virus and Virus-like Diseases of Major Crops in Developing Countries, Springer, Dordrecht, 2003.

[30] S.D.W. Frost, O.G. Pybus, J.R. Gog, C. Viboud, S. Bonhoeffer, T. Bedford, Eight challenges in phylodynamic inference, Epidemics 10 (2015) 88–92.

[31] P.T. Dolan, Z.J. Whitfield, R. Andino, Mapping the evolutionary potential of RNA viruses, Cell Host Microbe 23 (4) (2018) 435–446.

[32] J.L. Geoghegan, E.C. Holmes, The phylogenomics of evolving virus virulence, Nat. Rev. Genet. 19 (12) (2018) 756–769.

[33] A. Stern, M.T. Yeh, T. Zinger, M. Smith, C. Wright, G. Ling, R. Nielsen, A. Macadam, R. Andino, The evolutionary pathway to virulence of an RNA virus, Cell 169 (1) (2017) 35–46.e19.

[34] B.M. Muhire, A. Varsani, D.P. Martin, SDT: a virus vlassification tool based on pairwise sequence alignment and identity calculation, PLoS One 9 (9) (2014) e108277.

[35] R.C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Res. 32 (2004) 1792–1797.

[36] L.-T. Nguyen, H.A. Schmidt, A. von Haeseler, B.Q. Minh, IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies, Mol. Biol. Evol. 32 (1) (2014) 268–274.

[37] S. Kalyaanamoorthy, B.Q. Minh, T.K.F. Wong, A. von Haeseler, L.S. Jermiin, ModelFinder: fast model selection for accurate phylogenetic estimates, Nat. Methods 14 (6) (2017) 587–589.

[38] K. Katoh, K. Misawa, K. Kuma, T. Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, Nucleic Acids Res. 30 (14) (2002) 3059–3066.

[39] I. Letunic, P. Bork, Interactive Tree of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation, Nucleic Acids Res. 49 (W1) (2021) W293–w296.

[40] W. Sagemann, D.E. Lesemann, H.L. Paul, D. Adomako, G.K. Owusu, Detection and comparison of some Ghanaian isolates of cacao swollen shoot virus (CSSV) by enzyme-linked immunosorbent assay (ELISA) and immunoelectron microscopy (IEM) using an antiserum to CSSV strain 1A, J. Phytopathol. 114 (1) (1985) 79–89.

[41] R.C. Team, R: A Language and Environment for Statistical Computing, 2021. https://www.r-project.org/.

[42] L.J. Harmon, J.B. Losos, T.J. Davies, R.G. Gillespie, J.L. Gittleman, W.B. Jennings, K.H. Kozak, M.A. McPeek, F. Moreno-Roark, T.J. Near, A. Purvis, R.E. Ricklefs, D. Schluter, J.A. Schulte II, O. Seehausen, B.L. Sidlauskas, O. Torres-Carvajal, J. T. Weir, A.Ø. Mooers, Early bursts of body size and shape evoluation are rare in comparative data, Evolution 64 (8) (2010) 2385–2396.

[43] M. Pagel, Inferring the historical patterns of biological evolution, Nature 401 (6756) (1999) 877–884.

[44] M. Pagel, The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies, Syst. Biol. 48 (3) (1999) 612–622.

[45] M. Pagel, Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters, Proc. Roy. Soc. Lond. B Biol. Sci. 255 (1342) (1994) 37–45.

[46] K.P. Burnham, D.R. Anderson, Model Selection and Multimodel Inference: a Practical Information-Theoretic Approach, 2 nd ed ed., Springer-Verlag, New York, NY, 2002.

[47] E. Paradis, K. Schliep, Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R, Bioinformatics 35 (3) (2019) 526–528.

[48] Z. Yang, Computational Molecular Evolution, Oxford University Press, 2006.

[49] S.A. Ishikawa, A. Zhukova, W. Iwasaki, O. Gascuel, A fast likelihood method to reconstruct and visualize ancestral scenarios, Mol. Biol. Evol. 36 (9) (2019) 2069–2085.

[50] A. Rambaut, T.T. Lam, L. Max Carvalho, O.G. Pybus, Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen), Virus Evol 2 (1) (2016) vew007.

[51] M. Navascués, F. Depaulis, B.C. Emerson, Combining contemporary and ancient DNA in population genetic and phylogeographical studies, Mol Ecol Resour 10 (5) (2010) 760–772.

[52] G.G.R. Murray, F. Wang, E.M. Harrison, G.K. Paterson, A.E. Mather, S.R. Harris, M.A. Holmes, A. Rambaut, J.J. Welch, The effect of genetic structure on molecular dating and tests for temporal signal, Methods Ecol. Evol. 7 (1) (2016) 80–89.

[53] E.M. Volz, S.D.W. Frost, Scalable relaxed clock phylogenetic dating, Virus Evol 3 (2) (2017).

[54] S. Kumar, G. Stecher, M. Li, C. Knyaz, K. Tamura, Mega X: molecular evolutionary genetics analysis across computing platforms, Mol. Biol. Evol. 35 (6) (2018) 1547–1549.

[55] K. Tamura, F.U. Battistuzzi, P. Billing-Ross, O. Murillo, A. Filipski, S. Kumar, Estimating divergence times in large molecular phylogenies, Proc. Natl. Acad. Sci. USA 109 (47) (2012) 19333–19338.

[56] K. Tamura, Q. Tao, S. Kumar, Theoretical foundation of the RelTime method for estimating divergence times from variable evolutionary rates, Mol. Biol. Evol. 35 (7) (2018) 1770–1782.

[57] E. Obok, A. Wetten, J. Allainguillaume, Molecular evidence of cacao swollen shoot virus acquisition and retention by *Planococcus citri* (Risso) and P*seudococcus longispinus* (Targioni-Tozzetti) and *Pseudococcus viburni* (Signoret) mealybugs, International Journal of Scientific and Research Publications (IJSRP) 8 (3) (2018) 2250–3153.
[58] K. Kouakou, Diversité moléculaire du CSSV (Cocoa swollen shoot virus) et épidémiologie de la maladie du swollen shoot du cacaoyer (Theobroma cacao L.) en Côte d'Ivoire Université Félix Houphouët Boigny, Abidjan (2014) 141.
[59] P. Carrasco, F. de la Iglesia, S.F. Elena, Distribution of fitness and virulence effects caused by single-nucleotide substitutions in *Tobacco Etch Virus*, J. Virol. 81 (23) (2007) 12979–12984.
[60] W. van Rheenen, W.J. Peyrot, A.J. Schork, S.H. Lee, N.R. Wray, Genetic correlations of polygenic disease traits: from theory to practice, Nat. Rev. Genet. 20 (10) (2019) 567–581.
[61] P.H. Thrall, J.J. Burdon, Evolution of virulence in a plant host-pathogen metapopulation, Science 299 (5613) (2003) 1735–1737.
[62] M. Vignuzzi, C.B. López, Defective viral genomes are key drivers of the virus–host interaction, Nature Microbiology 4 (7) (2019) 1075–1087.
[63] R. Acosta-Leal, S. Duffy, Z. Xiong, R.W. Hammond, S.F. Elena, Advances in plant virus evolution: translating evolutionary insights into better disease management, Phytopathology 101 (10) (2011) 1136–1148.
[64] A. Addetia, Q. Phung, B.T. Bradley, M.J. Lin, H. Zhu, H. Xie, M.L. Huang, A. L. Greninger, In vivo generation of BK and JC polyomavirus defective viral genomes in human urine samples associated with higher viral loads, J. Virol. 95 (12) (2021) e00250-21.
[65] N. Chingandu, K. Kouakou, R. Aka, G. Ameyaw, O.A. Gutierrez, H.-W. Herrmann, J.K. Brown, The proposed new species, cacao red vein virus, and three previously recognized badnavirus species are associated with cacao swollen shoot disease, Virol. J. 14 (1) (2017) 199.
[66] M.J. Jeger, S.E. Seal, F. Van den Bosch, Evolutionary epidemiology of plant virus disease, Adv. Virus Res. 67 (2006) 163–203.
[67] P. Lefeuvre, D.P. Martin, S.F. Elena, D.N. Shepherd, P. Roumagnac, A. Varsani, Evolution and ecology of plant viruses, Nat. Rev. Microbiol. 17 (10) (2019) 632–644.
[68] D.P. Martin, J.A. Willment, R. Billharz, R. Velders, B. Odhiambo, J. Njuguna, D. James, E.P. Rybicki, Sequence diversity and virulence in Zea mays of Maize streak virus isolates, Virology 288 (2) (2001) 247–255.
[69] E. Domingo, J. Sheldon, C. Perales, Viral quasispecies evolution, Microbiol. Mol. Biol. Rev. 76 (2) (2012) 159–216.
[70] E. Domingo, C. Perales, Viral quasispecies, PLoS Genet. 15 (10) (2019) e1008271.
[71] M.B. Mandary, M. Masomian, C.L. Poh, Impact of RNA virus evolution on quasispecies formation and virulence, Int. J. Mol. Sci. 20 (18) (2019).
[72] J. Zeng, H. Wang, X. Xie, C. Li, G. Zhou, D. Yang, L. Yu, S. Perlman, Ribavirin-resistant variants of foot-and-mouth disease virus: the effect of restricted quasispecies diversity on viral virulence, J. Virol. 88 (8) (2014) 4008–4020.
[73] J.K. Pfeiffer, K. Kirkegaard, Increased fidelity reduces poliovirus fitness and virulence under selective pressure in mice, PLoS Pathog. 1 (2) (2005) e11-e11.
[74] C.E. Cressler, D.V. Mc Leod, C. Rozins, J. Van Den Hoogen, T. Day, The adaptive evolution of virulence: a review of theoretical predictions and empirical tests, Parasitology 143 (7) (2016) 915–930.
[75] G. Mangenot, H. Alibert, A. Basset, Sur les caractères du " Swollen-Shoot " en Côte-d'Ivoire, J. Agric. Tradit. Bot. Appl. 283–284 (1946) 173–184.
[76] G.J. Slater, M.W. Pennell, Robust regression and posterior predictive simulation increase power to detect early bursts of trait evolution, Syst. Biol. 63 (3) (2013) 293–308.
[77] G.A. Ameyaw, H.K. Dzahini-Obiatey, O. Domfeh, F.K. Oppong, K. Abaka-Ewusie, History and data analyses of 'cutting out' method for Cocoa swollen shoot virus disease (CSSVD) control in Ghana, J. Plant Dis. Prot. 122 (5/6) (2015) 200–206.
[78] S. Sacristán, F. García-Arenal, The evolution of virulence and pathogenicity in plant pathogen populations, Mol. Plant Pathol. 9 (3) (2008) 369–384.
[79] A. Fraile, F. García-Arenal, Environment and evolution modulate plant virus pathogenesis, Current Opinion in Virology 17 (2016) 50–56.
[80] F. Danquah, The Capsid, Black Pod and Swollen Shoot Cacao Diseases in Ghana, 1910-1966, Iowa State University, 1991.
[81] R. Darkoah, Diagnosis of Cocoa Swollen Shoot Virus Disease by Polymerase Chain Reaction (PCR), University of Ghana, 2000.
[82] H. Dzahini-Obiatey, R.T.V. Fox, Early signs of infection in cacao swollen shoot virus (CSSV) inoculated cocoa seeds and the discovery of the cotyledons of the resultant plants as rich sources of CSSV, Afr. J. Biotechnol. 9 (5) (2010) 593–603.
[83] A.F. Posnette, N.F. Robertson, J.M.A. Todd, Virus diseases of cacao in West Africa, Ann. Appl. Biol. 37 (2) (1950) 229–240.
[84] J. Thresh, T. Tinsley, in: F.P.P. Bulletin (Ed.), The Virus Complex Causing Swollen Shoot Disease of Cacao in West Africa, 1960, pp. 65–69.
[85] D. Dare, Studies on the Genetics of Cocoa Swollen Shoot Virus Disease (CSSVD) Resistance in Cocoa, University of Ghana, 1999.
[86] J. Hughes, L. Ollennu, The virobacterial agglutination test as a rapid means of detecting cocoa swollen shoot virus, Ann. Appl. Biol. 122 (2) (1993) 299–310.
[87] J.d.A. Hughes, L.A.A. Ollennu, Mild strain protection of cocoa in Ghana against cocoa swollen shoot virus—a review, Plant Pathol. 43 (3) (1994) 442–457.
[88] S. Sackey, Novel Technologies for Disease Indexing and Screening for CSSVD Resistance, International Workshop on New Technologies and Cocoa Breeding, INGENIC, 2000, pp. 155–162.
[89] K. Sastry, T. Zitter, Plant Virus and Viroid Diseases in the Tropics, first ed., Springer Dordrecht, 2013.
[90] E.C. Holmes, The evolutionary genetics of emerging viruses, Annu. Rev. Ecol. Evol. Systemat. 40 (1) (2009) 353–372.
[91] J. Thresh, The origins and epidemiology of some important plant virus diseases, Applied Biology 5 (1980) 1–65.
[92] A. Posnette, Transmission of swollen shoot, Trop. Agric. 17 (98) (1940).
[93] A. Chevalier, Un proche parent des Theobroma spontané en Côte-d'Ivoire et Gold Coast pays de grande culture du Cacaoyer, Revue Bot. Apple. Agric. Trop. 26 (1946) 304–306.
[94] J.M.A. Todd, An indigenous source of swollen shoot disease of cacao, Nature 167 (4258) (1951) 952–953.
[95] T.W. Tinsley, A.L. Wharton, Studies on the host ranges of viruses from *Theobroma cacao* L, Ann. Appl. Biol. 46 (1) (1958) 1–6.
[96] J.G. Bald, T.W. Tinsley, A quasi-genetic model for plant virus host ranges: IV. Cacao swollen-shoot and mottle-leaf viruses, Virology 40 (2) (1970) 369–378.
[97] E. Muller, Cacao swollen shoot virus (CSSV): history, biology, and genome, in: B. A. Bailey, L.W. Meinhardt (Eds.), Cacao Diseases: a History of Old Enemies and New Encounters, Springer International Publishing, Cham, 2016, pp. 337–358.
[98] T.W. Tinsley, The ecology of cacao viruses. I. The role of wild hosts in the incidence of swollen shoot virus in West Africa, J. Appl. Ecol. 8 (2) (1971) 491–495.
[99] J.C. Motamayor, P. Lachenaud, J.W. da Silva e Mota, R.G. Loor, D.N. Kuhn, J. S. Brown, R.J. Schnell, Geographic and genetic population differentiation of the Amazonian chocolate tree (*Theobroma cacao* L.), PLoS One 3 (10) (2008) e3311.
[100] E. Muller, I. Ullah, J.M. Dunwell, A.J. Daymond, M. Richardson, J. Allainguillaume, A. Wetten, Identification and distribution of novel badnaviral sequences integrated in the genome of cacao (*Theobroma cacao*), Sci. Rep. 11 (1) (2021) 8270.
[101] T. Frasch, The coming of cacao and chocolate to Ceylon, Food Hist. 12 (1) (2014) 137–152.
[102] R. Ciferri, F. Ciferri, The evolution of cultivated cacao, Evolution 11 (4) (1957) 381–397.
[103] A. Young, The Chocolate Tree: a Natural History of Cacao, first ed., Smithsonian Institution Press, 1994.
[104] I. Ullah, J.M. Dunwell, Bioinformatic, genetic and molecular analysis of several badnavirus sequences integrated in the genomes of diverse cocoa (*Theobroma cacao* L.) germplasm, Saudi J. Biol. Sci. 30 (5) (2023) 103648.
[105] R. Ramos-Sobrinho, M.M.M. Ferro, T. Nagata, A.S. Puig, C.V. Keith, D.S. Britto, O. A. Gutierrez, J.P. Marelli, J.K. Brown, Complete genome sequences of three newly discovered cacao mild mosaic virus isolates from *Theobroma cacao* L. in Brazil and Puerto Rico and evidence for recombination, Arch. Virol. 166 (7) (2021) 2027–2031.
[106] E.S.L. Santos, C.B.M. Cerqueira-Silva, G.M. Mori, D. Ahnert, D.L.N. Mello, J. L. Pires, R.X. Corrêa, A.P. Souza, Genetic structure and molecular diversity of cacao plants established as local varieties for more than two centuries: the genetic history of cacao plantations in Bahia, Brazil, PLoS One 10 (12) (2015) e0145276.
[107] I. Ullah, A.J. Daymond, P. Hadley, M.J. End, P. Umaharan, J.M. Dunwell, Identification of cacao mild mosaic virus (CaMMV) and cacao yellow vein-banding virus (CYVBV) in cocoa (theobroma cacao) germplasm, Viruses 13 (11) (2021).
[108] P.O. Aikpokpodion, J.C. Motamayor, V.O. Adetimirin, Y. Adu-Ampomah, I. Ingelbrecht, A.B. Eskes, R.J. Schnell, M. Kolesnikova-Allen, Genetic diversity assessment of sub-samples of cacao, *Theobroma cacao* L. collections in West Africa using simple sequence repeats marker, Tree Genet. Genomes 5 (4) (2009) 699–711.
[109] W.G. Clarence-Smith, Cocoa Pioneer Fronts since 1800. The Role of Smallholders, Planters and Merchants, first ed., Palgrave Macmillan, UK, 1996.
[110] T. Walker, Slave labor and chocolate in Brazil: the culture of cacao plantations in Amazonia and Bahia (17th–19th Centuries) 1, Food Foodw. 15 (1–2) (2007) 75–106.
[111] D. Oppong, Rationally Motivated Cocoa Production in Ghana: Motivations and De-motivations of Small-Scale Cocoa Producers in Fawohoyeden, University of Oslo, Ghana, 2014.
[112] K.P. Nair, Tree Crops. Harvesting Cash from the World's Important Cash Crops, Springer Cham, 2021.
[113] W.G. Clarence-Smith, African and European cocoa producers on Fernando Póo, 1880s to 1910s, J. Afr. Hist. 35 (2) (1994) 179–199.
[114] G. Lockwood, A comparison of the growth and yield during a 20 year period of Amelonado and upper amazon hybrid cocoa in Ghana, Euphytica 25 (1) (1976) 647–658.
[115] R. Ploetz, The impact of diseases on cacao production: a global overview, in: B. A. Bailey, L.W. Meinhardt (Eds.), Cacao Diseases. A History of Old Enemies and New Encounters, Springer Cham, 2016, pp. 33–59.
[116] I. Pagán, P. González-Jara, A. Moreno-Letelier, M. Rodelo-Urrego, A. Fraile, D. Piñero, F. García-Arenal, Effect of biodiversity changes in disease risk: exploring disease emergence in a plant-virus system, PLoS Pathog. 8 (7) (2012) e1002796.
[117] R. González, A. Butković, S.F. Elena, Role of host genetic diversity for susceptibility-to-infection in the evolution of virulence of a plant virus, Virus Evol 5 (2) (2019) vez024.
[118] E.H. Stukenbrock, B.A. McDonald, The origins of plant pathogens in agro-ecosystems, Annu. Rev. Phytopathol. 46 (2008) 75–100.
[119] S. Sacristán, E.M. Goss, S. Eves-van den Akker, How do pathogens evolve novel virulence activities? Mol. Plant Microbe Interact. 34 (6) (2021) 576–586.
[120] R.A.C. Jones, Plant virus emergence and evolution: origins, new encounter scenarios, factors driving emergence, effects of changing world conditions, and prospects for control, Virus Res. 141 (2) (2009) 113–130.
[121] C. Andres, W.J. Blaser, H.K. Dzahini-Obiatey, G.A. Ameyaw, O.K. Domfeh, M. A. Awiagah, A. Gattinger, M. Schneider, S.K. Offei, J. Six, Agroforestry systems

can mitigate the severity of cocoa swollen shoot virus disease, Agric. Ecosyst. Environ. 252 (2018) 83–92.
[122] J.R. Díaz-Valderrama, S.T. Leiva-Espinoza, M.C. Aime, The history of cacao and its diseases in the Americas, Phytopathology 110 (10) (2020) 1604–1619.
[123] J. Rojas-López, La producción de cacao en la Venezuela de la segunda mitad del siglo x viii: ¿grandes o modestas plantaciones? Revista Derecho y Reforma Agraria 0 (38) (2012) 88–109.
[124] I. Ullah, M. Kamran, J.M. Dunwell, Identification of a novel polerovirus in cocoa (*Theobroma cacao*) germplasm and development of molecular methods for use in diagnostics, Pathogens 12 (11) (2023) 1284.
[125] E.E. Obok, Cacao Swollen Shoot Virus in Nigeria: Analysis of a Pathogen and its Vectors, School of Agriculture, Policy and Development, University of Reading, 2015.
[126] A. Wetten, C. Campbell, J. Allainguillaume, High-resolution melt and morphological analyses of mealybugs (Hemiptera: pseudococcidae) from cacao: tools for the control of Cacao swollen shoot virus spread, Pest Manag. Sci. 72 (3) (2016) 527–533.
[127] M. Palma-Jiménez, M. Blanco-Meneses, C. Guillén-Sánchez, Las cochinillas harinosas (Hemiptera: pseudococcidae) y su impacto en el cultivo de Musáceas, Agron. Mesoam. 30 (1) (2019) 281–298.
[128] T.W. Kirkpatrick, Insect transmission of cacao virus disease in Trinidad, Bull. Entomol. Res. 41 (1) (2009) 99–117.
[129] A.S. Puig, S. Wurzel, S. Suarez, J.P. Marelli, J. Niogret, Mealybug (Hemiptera: pseudococcidae) species associated with cacao mild mosaic virus and evidence of virus acquisition, Insects 12 (11) (2021).
[130] K.S. Sastry, B. Mandal, J. Hammond, S.W. Scott, R.W. Briddon, Encyclopedia of Plant Viruses and Viroids, 1 ed., Springer New Delhi, 2019.
[131] E. Obok, A. Wetten, J. Allainguillaume, Electropenetrography application and molecular-based virus detection in mealybug (Hemiptera: pseudococcidae) vectors of cacao swollen shoot virus on *Theobroma cacao* L, Ann. Agric. Sci. 63 (1) (2018) 55–65.
[132] A.F. Posnette, N.F. Robertson, Virus diseases of cacao in West Africa, Ann. Appl. Biol. 37 (3) (1950) 363–377.
[133] P.F. Entwistle, J.F. Longworth, The relationships between cacao viruses and their vectors: the feeding behaviowr of three mealybug (Hornoptera: pseudococcidae) species, Ann. Appl. Biol. 52 (3) (1963) 387–391.
[134] O. Roivainen, Transmission of cocoa viruses by mealybugs (Homoptera: pseudococcidae), Agric. Food Sci. 48 (3) (1976) 203–304.
[135] C.A.M. Campbell, The influence of overhead shade and fertilizers on the Homoptera of mature Upper-Amazon cocoa trees in Ghana, Bull. Entomol. Res. 74 (1) (1984) 163–174.
[136] M. Bigger, The relative abundance of the mealybug vectors (Hemiptera: coccidae and Pseudococcidae) of Cocoa swollen shoot disease in Ghana, Bull. Entomol. Res. 71 (3) (1981) 435–448.
[137] M. Bousalem, E.J. Douzery, S.E. Seal, Taxonomy, molecular phylogeny and evolution of plant reverse transcribing viruses (family Caulimoviridae) inferred from full-length genome and reverse transcriptase sequences, Arch. Virol. 153 (6) (2008) 1085–1102.
[138] S.K. Sharma, P. Vignesh Kumar, A.S. Geetanjali, K.B. Pun, V.K. Baranwal, Subpopulation level variation of banana streak viruses in India and common evolution of banana and sugarcane badnaviruses, Virus Gene. 50 (3) (2015) 450–465.
[139] S. Duffy, L.A. Shackelton, E.C. Holmes, Rates of evolutionary change in viruses: patterns and determinants, Nat. Rev. Genet. 9 (4) (2008) 267–276.
[140] R. Sanjuán, M.R. Nebot, N. Chirico, L.M. Mansky, R. Belshaw, Viral mutation rates, J. Virol. 84 (19) (2010) 9733–9748.
[141] R. Sanjuán, P. Domingo-Calap, Mechanisms of viral mutation, Cell. Mol. Life Sci. 73 (23) (2016) 4433–4448.
[142] C. Firth, A. Kitchen, B. Shapiro, M.A. Suchard, E.C. Holmes, A. Rambaut, Using time-structured data to estimate evolutionary rates of double-stranded DNA viruses, Mol. Biol. Evol. 27 (9) (2010) 2038–2051.
[143] L.S. Hagen, M. Jacquemond, A. Lepingle, H. Lot, M. Tepfer, Nucleotide sequence and genomic organization of cacao swollen shoot virus, Virology 196 (2) (1993) 619–628.