JOURNAL OF
GENERAL VIROLOGY

RESEARCH ARTICLE
Mar *et al*., *Journal of General Virology* 2017;98:1537–1551
DOI 10.1099/jgv.0.000784

MICROBIOLOGY SOCIETY

Genetic variability and population structure of the New World begomovirus *Euphorbia yellow mosaic virus*

Talita Bernardon Mar,[1,2] César Augusto Diniz Xavier,[1,2] Alison Talis Martins Lima,[3] Angélica Maria Nogueira,[1,2] José Cleydson Ferreira Silva,[2] Roberto Ramos-Sobrinho,[4] Douglas Lau[5] and F. Murilo Zerbini[1,2,*]

Abstract

The emergence of begomoviruses (whitefly-transmitted viruses classified in the genus *Begomovirus*, family *Geminiviridae*) in Brazil probably occurred by horizontal transfer from non-cultivated plants after the introduction of *Bemisia tabaci* MEAM1. The centre of diversity of *Euphorbia heterophylla* (Euphorbiaceae) is located in Brazil and Paraguay, where it is an invasive species in soybean and other crops. Reports of possible begomovirus infection of *E. heterophylla* in Brazil date back to the 1950s. In 2011, *Euphorbia yellow mosaic virus* (EuYMV) was described in symptomatic plants collected in the Brazilian state of Goiás. Here we assess the genetic variability and population structure of begomoviruses infecting *E. heterophylla* in samples collected throughout nine Brazilian states from 2009 to 2014. A total of 158 and 57 haplotypes were compared in DNA-A and DNA-B datasets, respectively. Analysis comparing population structure in a large sampled area enabled us to differentiate two subpopulations. Further, the application of discriminant analysis of principal components allowed the differentiation of six subpopulations according to sampling locations and in agreement with phylogenetic analysis. In general, negative selection was predominant in all six subpopulations. Interestingly, we were able to reconstruct the phylogeny based on the information from the 23 sites that contributed most to the geographical structure proposed, demonstrating that these polymorphisms hold supporting information to discriminate between subpopulations. These sites were mapped in the genome and compared at the level of amino acid changes, providing insights into how genetic drift and selection contribute to maintain the patterns of begomovirus population variability from a geographical structuring point of view.

INTRODUCTION

The *Geminiviridae* family includes viruses whose genomes are composed of one or two molecules of circular, single-stranded DNA encapsidated by a single structural protein into twinned, quasi-icosahedral particles. The family includes the genera *Begomovirus, Becurtovirus, Capulavirus, Curtovirus, Eragrovirus, Grablovirus*, *Mastrevirus*, *Topocuvirus* and *Turncurtovirus*, based on insect vector type, host range, genome organization and phylogenetic relationships [1]. Begomoviruses are transmitted by the whitefly *Bemisia tabaci* and infect dicot plants. The genus *Begomovirus* can be divided into 'Old World' (OW; Europe, Africa, Asia and Australasia) and 'New World' (NW; the Americas) lineages, based on genome organization, phylogenetic relationships and geographical distribution [2, 3]. Most NW begomoviruses have two components, named DNA-A and DNA-B. The DNA-A contains genes involved in replication and encapsidation of the viral progeny. The DNA-B contains genes required for intra- and intercellular movement in the plant. Both genomic components are required for systemic infection of the host [3].

Geminivirus populations, including begomoviruses, possess high genetic variability, mostly due to their high nucleotide substitution rates, which are similar to those estimated for RNA viruses (of the order of 10^{-4} substitutions per site per year) [4, 5], and the frequent occurrence of recombination [6] and pseudorecombination between bipartite viruses [7]. The emergence of begomoviruses in Brazil probably occurred through horizontal transfer of indigenous viruses infecting non-cultivated plants after the introduction of *Bemisia tabaci* Middle East–Asia Minor 1 (MEAM1; previously known as

Received 25 November 2016; Accepted 20 March 2017
Author affiliations: [1]Dep de Fitopatologia/BIOAGRO, Universidade Federal de Viçosa, Viçosa, MG 36570-900, Brazil; [2]National Research Institute for Plant–Pest Interactions, Universidade Federal de Viçosa, Viçosa, MG 36570-900, Brazil; [3]Instituto de Ciências Agrárias, Universidade Federal de Uberlândia, Uberlândia, MG 38400-902, Brazil; [4]Centro de Ciências Agrárias/Fitossanidade, Universidade Federal de Alagoas, Rio Largo, AL 57100-000, Brazil; [5]Embrapa Trigo, Rodovia BR-285, CP 3081, Passo Fundo, RS 99001-970, Brazil.
***Correspondence:** F. Murilo Zerbini, zerbini@ufv.br
Keywords: begomovirus; geminivirus; genetic variability; non-cultivated hosts; population structure.
Abbreviations: AMOVA, analysis of molecular variance; DAPC, discriminant analysis of principal components; EuYMV, *Euphorbia yellow mosaic virus*; MEAM1, Middle East-Asia Minor 1; NW, New World; OW, Old World; RDP4, Recombination Detection Program version 4.
The sequences reported in this paper have been deposited in GenBank under access numbers KY559430–KY559646.
Three supplementary figures and eight supplementary tables are available with the online Supplementary Material.

B. tabaci biotype B) in the mid-1990s [8–10]. Since the introduction of *B. tabaci* MEAM1, a large number of begomoviruses have been described in association with tomato and with many non-cultivated plants [11–15]. The presence of several viruses transmitted by the same vector to a wide range of hosts facilitates mixed infections, where novel recombinant variants with increased fitness can be formed [16–18]. Evidence for the emergence of tomato-infecting begomoviruses in Brazil, involving both recombination and pseudorecombination processes, has been reported by Silva *et al.* [13].

Euphorbia heterophylla is an invasive species in soybean and other crops in Brazil and Paraguay, its centre of origin [19]. Its occurrence is common in the southern, southeastern and mid-western regions of Brazil [20]. There are reports of *E. heterophylla* herbicide-resistant genotypes in the state of Rio Grande do Sul, where it is present in 74 % of soybean-producing areas [21, 22]. Reports of possible begomoviruses infecting *E. heterophylla* in Brazil date back to the 1950s [23]. In 2011, *Euphorbia yellow mosaic virus* (EuYMV) was reported in *E. heterophylla* plants collected in the state of Goiás [24]. Later, the virus was found in Brazil in the non-cultivated plant species *Sida santaremnensis* [25], *Macroptilium atropurpureus* [12] *and Crotalaria juncea* [26]. Experimentally, *Arabidopsis thaliana* is also a host of the virus, in which it was demonstrated to enhance recombination rates [27]. Barreto *et al.* [26] demonstrated that *E. heterophylla* is one of the hosts of *Tomato severe rugose virus* (ToSRV), although with a low viral titre. Tests of free choice, adult preference and oviposition indicated that *E. heterophylla* was the most suitable host for *Bemisia tabaci* MEAM1 among seven non-cultivated plants tested [28].

Several non-cultivated plants have been reported as begomoviruses hosts. These plants may serve as virus reservoirs, from which they can be transmitted to cultivated plants, and also as 'mixing vessels' that increase the probability of interspecific recombination and pseudorecombination [13, 26, 29]. Given the high preference of whiteflies for *E. heterophylla*, the possibility of co-infection with ToSRV (and possibly other crop-infecting begomoviruses), and the effects of EuYMV on the host's recombination rate, the study of EuYMV populations in this non-cultivated plant could provide valuable clues regarding the probability of it infecting crop plants due to its relatively high rate of evolution.

We assessed the begomovirus diversity associated with *E. heterophylla* plants, and estimated the genetic variability of EuYMV populations infecting this host. The genetic structure of EuYMV populations was investigated in a large-scale sampling, with isolates collected in locations throughout Brazil from 2009 to 2014, using Bayesian clustering and multivariate statistical analyses. We also highlighted the genome sites that contributed most to the geographical structure proposed and the processes that maintained genetic structure, and described how the genetic variability is distributed within EuYMV subpopulations.

RESULTS

Viral detection and recombination analysis

We investigated 165 symptomatic *E. heterophylla* samples, collected over a period of 6 years in locations throughout Brazil (states of Amazonas, Goiás, Mato Grosso do Sul, Minas Gerais, Paraná, Pernambuco, Rio Grande do Sul and Santa Catarina, and the Federal District). Based on the detection of a 2600 bp band after digestion of the rolling-circle amplification (RCA) products with restriction enzymes, 142 samples were preliminarily positive for the presence of a begomovirus (data not shown; also, over the years we have tested symptomless samples and found them to be consistently virus-free). Full-length DNA-A and DNA-B components were cloned from 129 and 41 RCA-positive samples, respectively, for a total of 133 samples from which at least one viral DNA component was cloned. EuYMV was the only virus detected in these samples, except for one sample which was infected by a virus belonging to a new species (to be reported elsewhere). Based on BLAST analysis and pairwise sequence comparisons we identified 150 full-length DNA-A (139 haplotypes) and 57 full-length DNA-B (52 haplotypes) components with >96 and >94 % identity amongst themselves and with EuYMV sequences retrieved from GenBank (data not shown). The EuYMV sequences determined in this study were combined with those retrieved from GenBank for a total of 158 and 57 haplotypes in the DNA-A and DNA-B datasets, respectively (Tables S1 and S2, available in the online Supplementary Material).

As expected, the DNA-B dataset was more variable [30], even with a smaller number of isolates. Indeed, the average number of nucleotide (nt) differences between isolates was 62.8 and 101.8 for the DNA-A and DNA-B, respectively, and the average pairwise number of nt differences was 0.02 and 0.04 for the DNA-A and DNA-B, respectively (Table 1).

No recombination events were detected in the intraspecific DNA-A dataset. In contrast, five recombinant isolates were detected in the intraspecific DNA-B dataset (Table S3). Events 1 and 3 involved breakpoints located within the *NSP* and/or *MP* genes, while event 2 presented breakpoints in non-coding regions. The topological differences between trees based on nt sequences of the *NSP* and *MP* genes further supported the recombinant origin of these isolates (Fig. S1b). Although recombination events were not assigned to long branches as expected [31], this was probably due to the high degree of similarity between recombinants and parental sequences.

Phylogenetic analysis

Bayesian-inferred trees based on nt sequences of the *CP*, *Rep*, *MP* and *NSP* genes and on the full-length DNA-A and DNA-B clearly clustered EuYMV isolates according to sampling locations (Figs 1 and S1). Both DNA-A and DNA-B trees showed two major clades supporting the existence of different subpopulations.

Table 1. Genetic variability of *Euphorbia yellow mosaic virus* (EuYMV)

Population	No. of sequences	Genome size (nt)	s	Eta	k	π	h	Hd
DNA-A								
Phylogeny/STRUCTURE								
Cluster I	68	2595	434	491	56.2	0.022	66	0.99
Cluster II	90	2601	517	605	48.3	0.019	89	1.00
DAPC								
pop1	73	2612	452	520	40.8	0.016	73	1.00
pop2	17	2606	159	167	40.5	0.016	16	0.99
pop3	7	2606	144	154	55.7	0.021	7	1.00
pop4	21	2603	161	170	27.4	0.011	19	0.99
pop5	31	2633	246	266	33.2	0.013	31	1.00
pop6	9	2608	48	48	11.0	0.004	9	1.00
Total	158	2589	699	865	62.8	0.024	154	1.00
DNA-B								
Cluster I	21	2558	462	528	93.1	0.036	20	1.00
Cluster II	36	2554	560	661	84.9	0.033	36	1.00
Total	57	2635	739	925	101.8	0.040	56	0.99

s, total number of polymorphic segregating sites; Eta, total number of mutations; k, average number of nucleotide differences between sequences; π, nucleotide diversity; h, haplotype number; Hd, haplotype diversity; DAPC, discriminant analysis of principal components.

The DNA-A tree showed strong support for isolates from AM, GO/DF, MG, PB and PE (cluster I) comprising a separate subpopulation from those of PR, RS and SC (cluster II). Isolates from MS were considered to be in cluster I (see below). Cluster I is supported by high posterior probability values, except for isolate BR:5:LEA:08 in the DNA-A dataset. This isolate could be a recombinant, since it is related (in a poorly supported clade) to RS isolates in the *Rep* gene phylogeny (Fig. 1 and S1a). In addition, three well-supported subclusters, including isolates from PE, MG and GO/DF, indicate further geographical structuring. This was not observed in cluster II, where isolates from PR, RS and SC were mixed in well-supported subclades.

The majority of isolates from MS clustered together in a well-supported clade with isolate BR:Amp1068:11, which could be a migrant. However, BR:Mir1:07 and BR:Pop709:10 are in a branch containing isolates from PB in the DNA-A and *Rep* phylogenies (Figs 1 and S1a). Observing the incongruence between *Rep* and *CP* phylogenies, these isolates could be recombinants that are not detected by RDP4. The *Rep* phylogeny clustered isolates from MS in a well-supported clade with the isolates from PB and with 17 isolates from RS (Fig. S1a). Interestingly, these same 17 isolates from RS clustered together in the DNA-A tree.

Even with only a few isolates representing each group, the DNA-B dataset broadly supports the same clusters (Figs 1 and S1), except for isolates BR:Vic54:14 and BR:Sau937:11, which could be migrants, and the recombinant BR:Ats3.2:09. In the DNA-B tree, BR:Mir1:07 and BR:Mir2:07 clustered with PR and RS isolates.

Genetic structure and variability of the EuYMV population

The clusters inferred from the Bayesian trees suggested the existence of two different subpopulations according to sampling locations. To evaluate spatial processes driving population structure we compared genetic divergence with geographical distances between isolates using Mantel's test. A significant correlation (r=0.5698; P<0.001) between the number of sites that differ between each pair of sequences with the geographic distance between isolates reinforces the suggestion that subpopulations assigned to geographical regions exist (Fig. 2a). Further, the DNA-A and DNA-B of cluster I isolates both displayed slightly greater genetic variability than those of cluster II isolates (Table 1), with significant differences between the π values of 0.022 and 0.019 for the DNA-A of cluster I and cluster II isolates, respectively, and 0.036 and 0.033 for the DNA-B of cluster I and cluster II isolates, respectively (Table 1 and Fig. 2c). This is consistent with nucleotide diversity calculated on a sliding window across the DNA-A and DNA-B (Fig. 1). DNA-B clusters were more variable than DNA-A clusters (Fig. 2c).

Differences in genetic variability between the phylogenetically inferred clusters supported the idea of population subdivision, which was confirmed by Nst statistics (Table 2). Analysis of molecular variance (AMOVA) attributed 39 and 27 % of total variation in the DNA-A and DNA-B datasets, respectively, to the variation between cluster I and cluster II isolates (Table 3). AMOVA analysis attributing collection year as a component of subdivision was also performed, with the results indicating that this parameter explains 20 % of the variation among the isolates (Table S4).

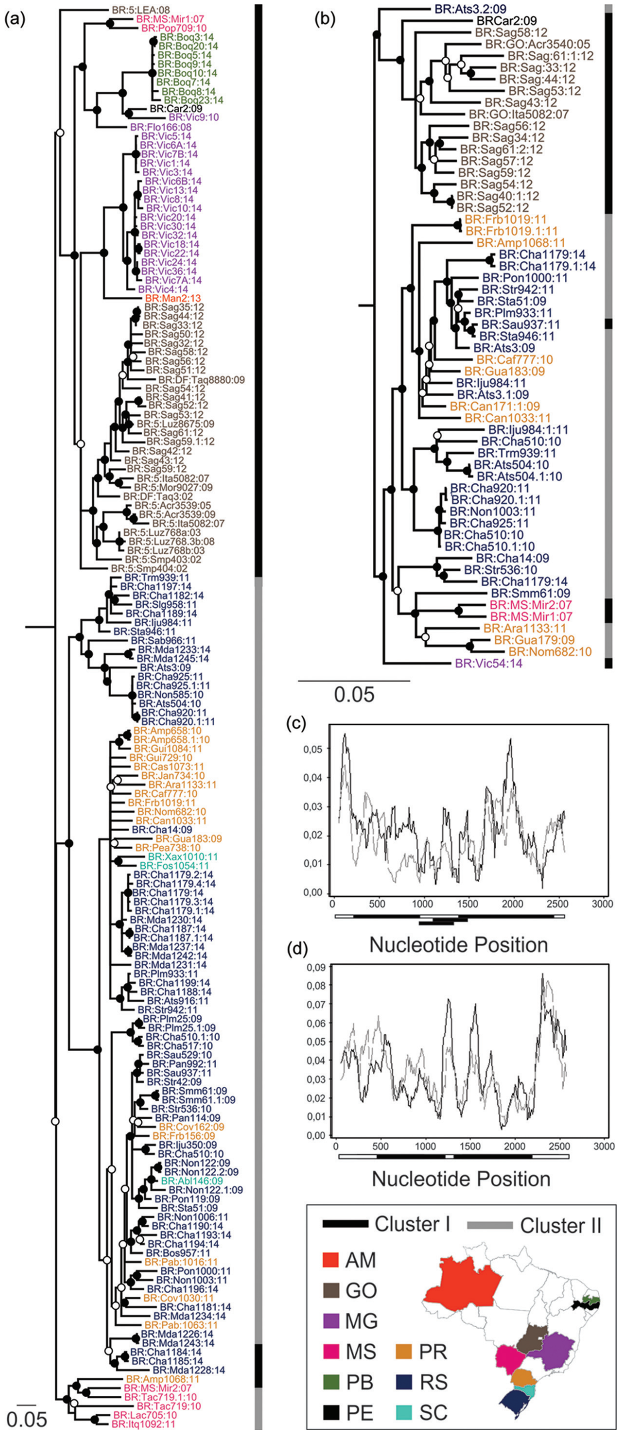


Fig. 1. Midpoint-rooted Bayesian inference trees based on full-length nucleotide sequences of EuYMV DNA-A (a) and DNA-B (b) components. Nodes to the right of branches with posterior probabilities equal to or higher than 0.8 are indicated by filled circles and those with values lower than 0.8 and higher than 0.5 are denoted by empty circles. The scale bars indicate nucleotide substitutions per site. Isolate colours based on geographical region of origin: AM, Amazonas; GO/DF, Goiás and the Federal District; MG, Minas Gerais; MS, Mato Grosso do Sul; PB, Paraíba; PE, Pernambuco; PR, Paraná; RS, Rio Grande do Sul; SC, Santa Catarina. Populations inferred by STRUCTURE are indicated by black/grey vertical bars. Mean pairwise number of nucleotide differences per site (nucleotide diversity, π) for the DNA-A (c) and DNA-B (d), calculated on a 10 nucleotide sliding window for isolates in cluster I (black lines) and cluster II (grey lines). A linearized representation of each DNA component, with genes indicated by black bars, is presented at the bottom of each graph.

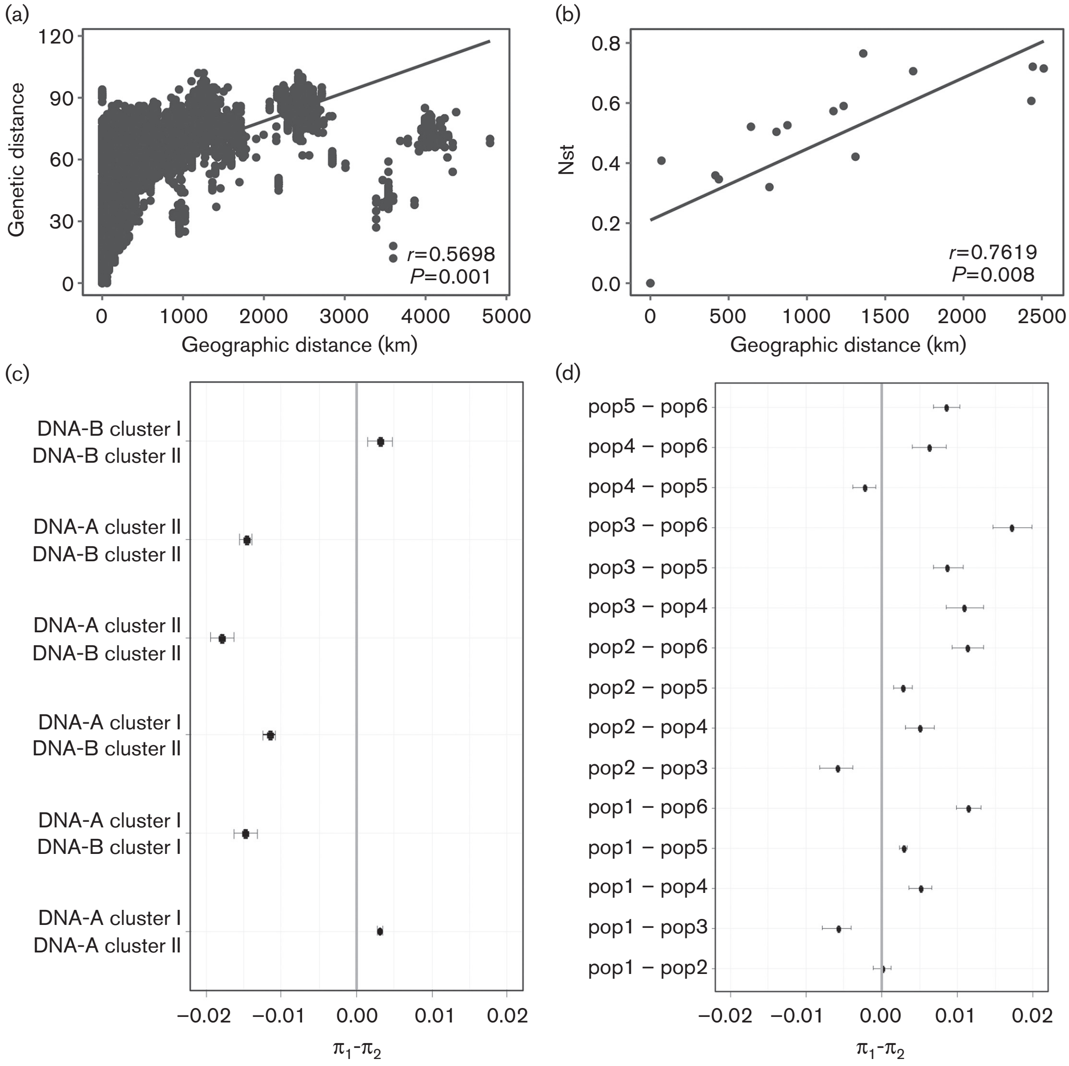


Fig. 2. Genetic divergence among geographical subpopulations of EuYMV. Mantel test correlating (a) genetic distance estimated by the number of sites that differ between each pair of DNA-A sequences with the geographical distance between isolates and (b) genetic distance between pop1 to pop6 expressed as pairwise Nst, with the geographical distance between subpopulations estimated by the mean geographical coordinates. Correlation coefficients and significances are indicated by *r* and *P*, respectively. Statistical significance of the differences amongst the average pairwise number of nucleotide differences per site (nucleotide diversity, π) calculated for (c) DNA-A and DNA-B Cluster I inferred by Bayesian trees and (d) subpopulations inferred by DAPC. Confidence intervals that include the value 'zero' denote no statistically significant difference between the means.

A Bayesian clustering method was also applied to infer distinct subpopulations. Due to the low evidence of recombination, we tested the degree of linkage equilibrium. Although we found evidence of significant linkage disequilibrium (LD; $P<0.001$), the corresponding standardized indexes of association (I^{s}_{A}) were 0.0147 and 0.0158 when only considering segregating sites, and 0.0045 and 0.0051 when considering the whole genome for the DNA-A and DNA-B datasets, respectively. These values were lower than those estimated for begomovirus datasets (0.0367) considered to be effectively in linkage equilibrium [32].

The results of STRUCTURE were consistent with phylogenetic analysis, and the EuYMV population was estimated to be composed of two subpopulations (clusters I and II; Fig. S2). In general, the genetic composition of each DNA-A component assigned into a cluster was not admixed, reflecting the absence of intraspecific recombination according to RDP analysis (Fig. S2). We detected only 27 out of 158 isolates with more than 10 % of admixing in their genetic composition. Seven of them, located in cluster I, were BR:5: LEA:08, the isolate from AM, and six isolates from MS. The MS isolates showed approximately 67 % genetic composition from cluster I, except for BR:Mir1:07 and BR:

Table 2. Results of subdivision tests performed on *Euphorbia yellow mosaic virus* (EuYMV) subpopulations

Population		Nst*
DNA-A		
Cluster I	Cluster II	0.323
pop1	pop2	0.408
pop1	pop3	0.359
pop1	pop4	0.573
pop1	pop5	0.504
pop1	pop6	0.721
pop2	pop3	0.346
pop2	pop4	0.590
pop2	pop5	0.526
pop2	pop6	0.715
pop3	pop4	0.421
pop3	pop5	0.320
pop3	pop6	0.607
pop4	pop5	0.521
pop4	pop6	0.765
pop5	pop6	0.706
DNA-B		
Cluster I	Cluster II	0.274

*Nst is an analogue of Wright's fixation index at the nucleotide sequence level [68, 69]; Values from 0 to 0.05 indicate little genetic differentiation between subpopulations; 0.05 to 0.15, moderate differentiation; 0.15 to 0.25, great differentiation; >0.25 high differentiation.

Pop709:10, which showed 99 and 86 % genetic composition from cluster I, respectively. Therefore, we considered these individuals to be members of cluster I instead of cluster II (as suggested by phylogenetic analysis). A similar pattern of admixed individuals from MS was identified in the DNA-B dataset (plus one sequence from MG; Fig. S2). The other 19 DNA-A admixed isolates, located in cluster II, included BR: Amp1068:11 and BR:Nom682:10, plus the same 17 isolates from RS, which were located in a well-supported clade with MS and PB isolates in the *Rep* phylogeny.

In the DNA-B dataset, nine admixed isolates were identified in cluster II: BR:Ara1133:11, the recombinant BR:Ats3.2:09 and seven individuals from PR and RS, which were clustered with MS isolates in the DNA-B phylogenetic tree. These results suggest that gene flow (between MS/PB and South isolates) and recombination contribute to increased genetic variability in these subpopulations.

Comparing the phylogenetic analysis and STRUCTURE results, we were able to explain the population structure based on two main groups. However, it is still remarkable that four well-supported subclusters of isolates from PB, MG, GO/DF and MS are present in the DNA-A tree (Fig. 1). Further, the year of collection also seems to contribute to the variability observed, although there is a significant overlap between locality and year of collection (Tables S1 and S2). Thus, to better understand the population structure and find the most suitable subdivision to explain the data, we performed a discriminant analysis of principal components (DAPC) on the same datasets. Since DAPC requires prior groups, we inferred genetic clusters using K-mean runs with a sequentially increasing number of K. The higher likelihood was obtained for six subpopulations (K=6). The subpopulations were checked for their association with geographical location. A correlation was found between the subpopulations inferred by K-means and the geographical distribution of the sampling locations (Fig. 3a). Isolates from GO/DF, MG, MS and PB were placed in different groups, and one group (group 4) comprised individuals from PR, SC and RS. Interestingly, the same 17 admixed isolates from RS were placed in a different group (group 5) (Fig. 3a). Accordingly, we proposed six subpopulations (pop1–6) based on geographical distribution (Fig. 3d) and checked how this fits in comparison with K-means (Fig. 3b). Most of the isolates placed in groups 1, 2, 3, 4, 5 and 6 based on geographical location (Fig. 3a, d) fall into pops 5, 3, 6, 1, 2 and 4, respectively (Fig. 3b, d). Interestingly, the detection of individuals from pop2 appears to have increased over the years (one detection in 2009, two in 2010, nine in 2011 and five in 2014), suggesting that this population may be expanding.

The ability of DAPC to discriminate between groups is heavily dependent on the number of principal components (PCs) retained. The trade-off between power of discrimination and over-fitting was measured by repeated DAPC analysis using randomized groups, computing a-scores for each group (Fig. S3c). For cross-validation analysis the dataset was divided into two sets: a training set and a validation set. DAPC was carried out in the training set with different numbers of PCs retained, and the ability to predict the membership of the validation set was used to confirm the number of PCs to be retained (Fig. S3d). We also checked the percentage of successful reassignment after randomization. Retaining 12 PCs allowed the six proposed subpopulations to be discriminated (Fig. 3c). Given that the number of investigated clusters was relatively low, all the discriminant functions were retained. DAPC classification was consistent with the proposed subpopulations, except for two discrepant isolates: BR: Amp1068:11 and BR:Vic9:10 were initially placed into pop1 and pop4, while DAPC classification assigned them to pop3 and pop6, respectively (Fig. S3a). We also checked the proper assignment of each isolate into a subpopulation by spatial analysis of molecular variance (SAMOVA), which considered populations that were geographically homogeneous and maximally differentiated from each other. SAMOVA attributed BR:Man2 : 13 to a unique subpopulation and pop1 and pop2 to the same group (Table S5).

The DAPC results were consistent with Nst population differentiation statistics (Table 2), and indicated genetic differentiation amongst isolates sampled from the geographical locations in Fig. 3(d). The correlation between Nst and geographic distance among subpopulations was also significant (r=0.7619; P<0.008) according to Mantel's test (Fig. 2b). Differences in genetic variability also supported the DAPC subdivision hypothesis (Table 1 and Fig. 2d). AMOVA

Table 3. Analysis of molecular variance (AMOVA) performed on *Euphorbia yellow mosaic virus* (EuYMV) subpopulations

Analysis	Source of variation	d.f.	Square sum	Variance components	% of variation
DNA-A					
Cluster I–II	Between populations	1	1444.006	18.26883 Va	38.85
	Within populations	156	4484.988	28.74992 Vb	61.15
	Total	157	5928.994	47.01875	
	Fst: 0.388				
DNA-B					
Cluster I–II	Between populations	1	556.697	19.08365 Va	27.43
	Within populations	55	2776.286	50.47792 Vb	72.57
	Total	56	3332.982	69.56157	
	Fst: 0.274				
DNA-A					
pop1–6	Among populations	5	2999.992	25.75366 Va	57.20
	Within populations	152	2929.002	19.26975 Vb	42.80
	Total	157	5928.994	45.02340	
	Fst: 0.572				

Fst, Wright's fixation index [68]; d.f., degrees of freedom.

analysis attributed 57 % of total variation to the variation among subpopulations (Table 3), which is significantly higher than the values obtained for clusters I and II and for year of collection (37 and 20 %, respectively; Tables 3 and S4). Hierarchical AMOVA, considering the STRUCTURE and DAPC results, attributed 19 % of the total variation to the variation between cluster I and II isolates, and 41 % to the variation among the six subpopulations (Table S4). Taking the number of isolates into account, pop3 is the most variable (seven isolates with π=0.021) (Table 1). Pop3 mostly comprises individuals from MS, where the agricultural landscape has a greater diversity of hosts.

It was not possible to extrapolate the same groups for the DNA-B dataset, except for pop5 (Fig. 3d), probably because there were not enough isolates representing each subpopulation. Four subpopulations explained the DNA-B dataset, and they reflected exactly the four main clades in the phylogenetic tree presented in Fig. 1 (data not shown).

Amino acid sites under selection

The kind of selection acting on the coding regions of each subpopulation was investigated by neutrality tests. All values were negative, and statistically significant deviations from neutrality were observed for the different genes, except for pop2 and pop3 (Table S6). These results indicate purifying selection acting on each population, or a recent population expansion. In agreement with previous studies [29, 31], all datasets showed d_N/d_S ratios lower than 1, except for the *Trap* gene in pop6, again indicating the predominance of purifying (negative) selection in all subpopulations. In fact, for all genes in all subpopulations, a higher number of sites under negative selection was detected (Table S7).

A wide variation of the d_N/d_S ratios for each gene/population indicated different selective constraints (Table S6). The d_N/d_S values for the *Trap* gene in all subpopulations and for the *CP* gene in pop6 were quite high, but even with a reduced number of negatively selected sites, the proportion of individual sites detected to be under negative selection was higher than under positive selection (Table S7). Interestingly, we found evidence of positive selection for the *Trap* gene in pop1 by PARRIS, and one positively selected site (codon 73) was detected by SLAC, FEL and IFEL. Also, evidence of positive selection was detected in *Rep* sites (codons 39 and 54 in pop5 and pop1, respectively) by four methods (SLAC, FEL, IFEL and REL; data not shown). Even with evidence of positive selection in these three sites, negative selection was shown to be predominant in all subpopulations.

Sites contributing to genetic divergence among subpopulations

We aimed to assess the 57 % variability among subpopulations (Table 3), identifying the genome sites that contributed most to the geographical structure defined by DAPC analysis (Fig. 3). Sites that presented substantial differences across subpopulations contribute substantially to the discrimination between subpopulations [33]. A plot with each site's contribution was used to assess the sites that best discriminate the proposed subpopulations, allowing us to point out the coding regions that drive genetic divergence among isolates from different sampling locations. A total of 23 sites reflected the geographical subdivision to the greatest degree, considering all discriminant functions (Fig. 4a). Different sites were identified by each discriminant function (DF), except sites #1254, #1259 (identified by DFs 3 and 5) and #1969 (identified by DFs 1 and 2). Since each DF corresponds to distinct levels of differentiation between each subpopulation, we considered all contributing sites in the analysis. Interestingly, we were able to discriminate between subpopulations based solely on the information from the sites that contributed most (Fig. 5). The information from these polymorphisms could be important for explaining the

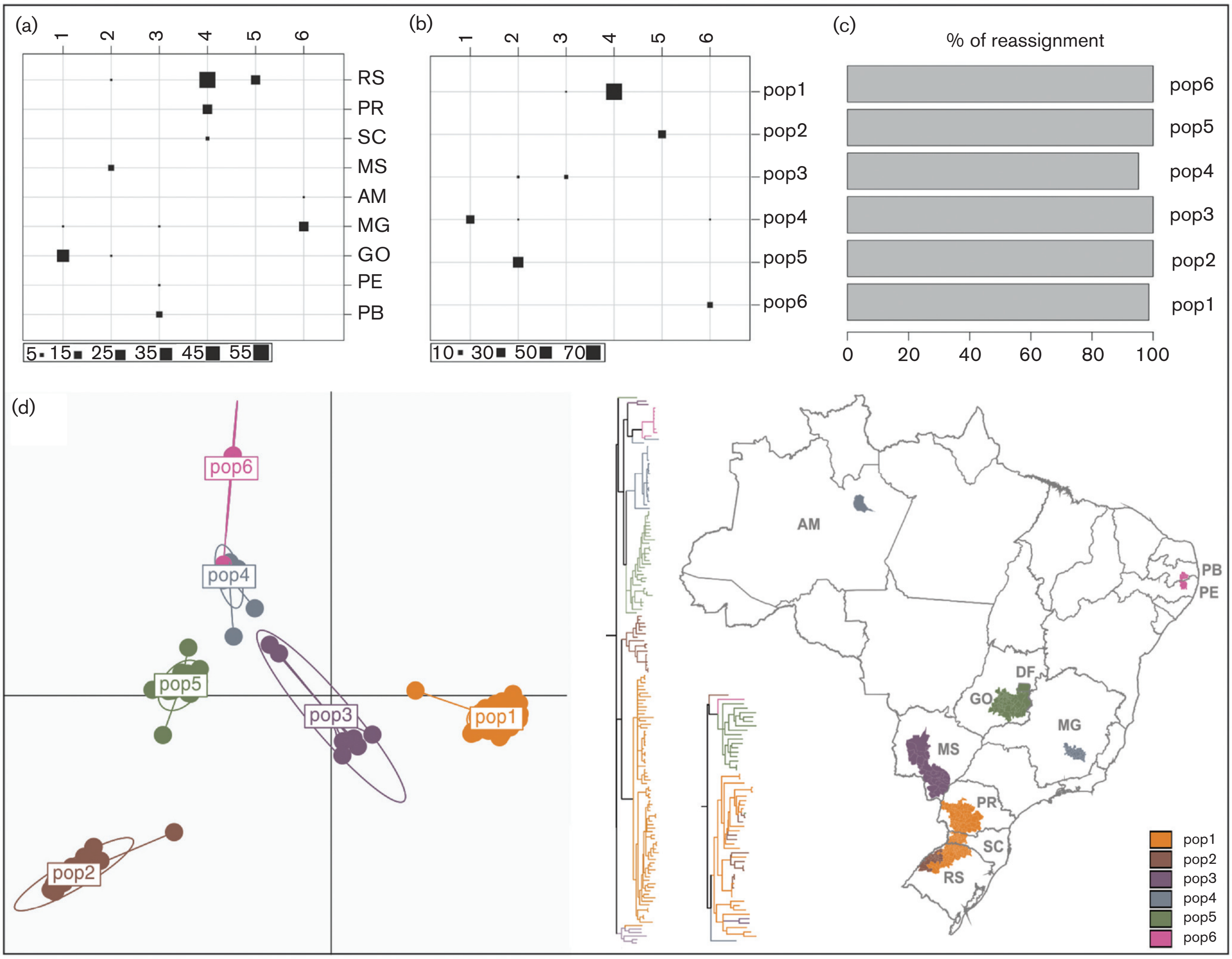
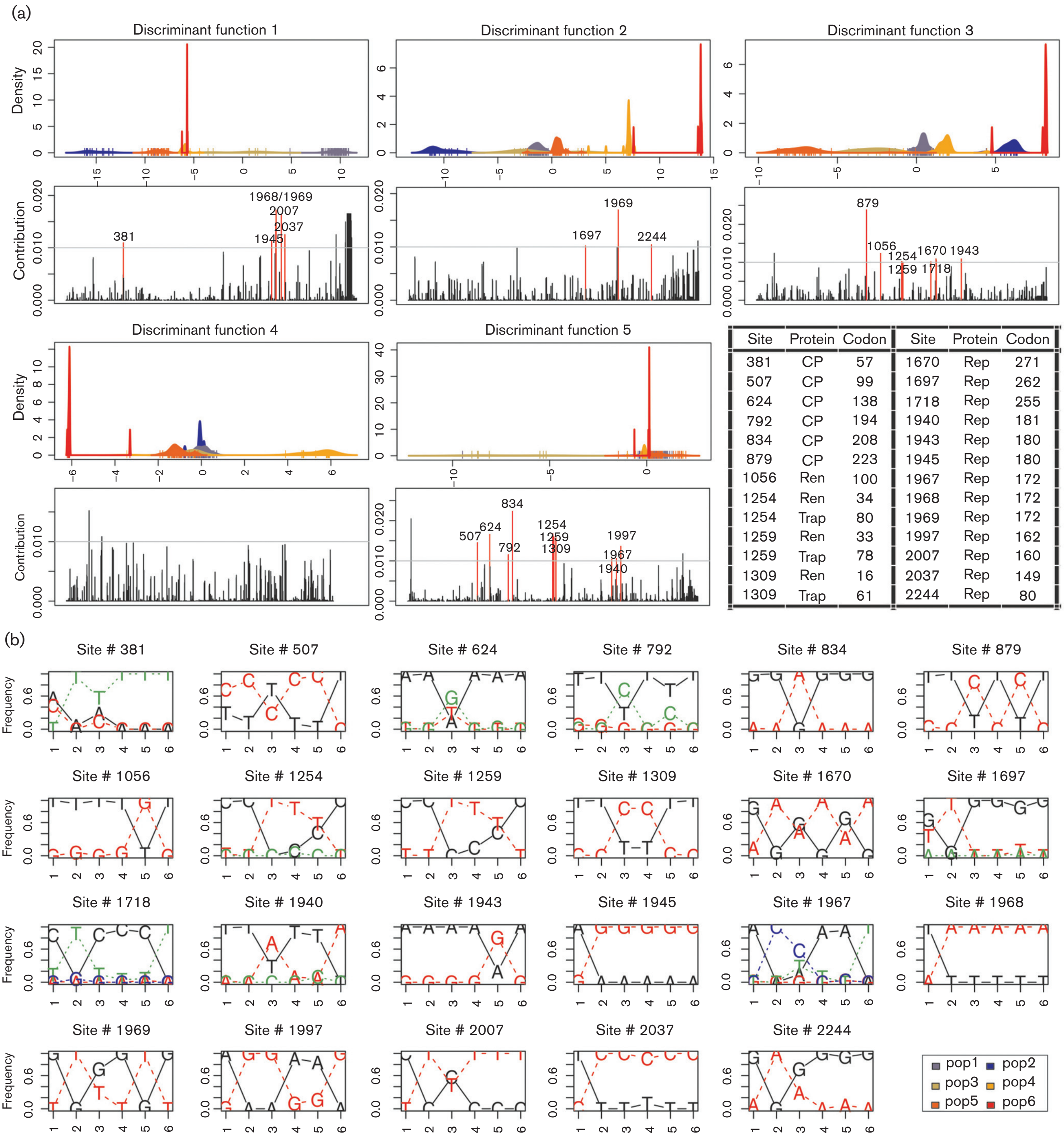


Fig. 3. Multivariate statistical clustering analysis of population subdivision using discriminant analysis of principal components (DAPC) for EuYMV DNA-A. (a) Comparison between groups inferred by *K*-means (columns) and sampling locations (rows: AM, Amazonas; GO/DF, Goiás and the Federal District; MG, Minas Gerais; MS, Mato Grosso do Sul; PB, Paraíba; PE, Pernambuco; PR, Paraná; RS, Rio Grande do Sul; SC, Santa Catarina.). (b) Comparison between groups inferred by *K*-means (columns) and geographical subdivision (rows). (c) Percentage of successful reassignment after randomization retaining 12 principal components. (d) DAPC scatterplot with ellipses representing subpopulations. Dots represent each isolate, coloured based on the geographical region located on the map and in the DNA-A (larger) and DNA-B (smaller) phylogenetic trees.

genetic differentiation between subpopulations, and may contain valuable clues to understand the main evolutionary mechanisms (selection and genetic drift) underlying the differentiation.

There were 12 sites covering the *CP* (sites 381, 507, 624, 792, 834 and 879), *Ren* (site 1254), *Trap* (site 1309) and *Rep* (site 1670, 1697, 1940 and 1997) coding regions displaying synonymous substitutions (Table S8) with different allele frequencies in each sampling location (Fig. 4b). For these sites the negative selection maintained the same amino acid being translated in all subpopulations (Table S7). However, we found different alleles either fixed in the third position of the codon or changing in frequency, yielding random fluctuations in the number of variants present in each subpopulation (Fig. 4b). For example, codon 208 encoded glutamic acid, but in the third position (site 834) the allele G was fixed in all subpopulations, except in pop3, which presented allele A (Fig. 4b). The mechanism underlying these changes could be genetic drift.

Sites 1254 and 1309, located in the overlapping *Trap* and *Ren* genes, were particularly informative. For example, site 1254 displayed a non-synonymous substitution under positive selection, corresponding to codon 80 in *Trap*, and a synonymous substitution encoding valine, corresponding to codon 34 of *Ren* (Table S8). Two possible residues can be translated in the *Trap* gene, hydrophobic glycine (allele C)



Site	Protein	Codon	Site	Protein	Codon
381	CP	57	1670	Rep	271
507	CP	99	1697	Rep	262
624	CP	138	1718	Rep	255
792	CP	194	1940	Rep	181
834	CP	208	1943	Rep	180
879	CP	223	1945	Rep	180
1056	Ren	100	1967	Rep	172
1254	Ren	34	1968	Rep	172
1254	Trap	80	1969	Rep	172
1259	Ren	33	1997	Rep	162
1259	Trap	78	2007	Rep	160
1309	Ren	16	2037	Rep	149
1309	Trap	61	2244	Rep	80

Fig. 4. Contribution of EuYMV DNA-A sites to the genetic divergence among geographical subpopulations. (a) Densities of isolates, coloured based on geographical subpopulation and the sites that contribute most, given a threshold of 0.01 in each discriminant function of DAPC. The height of each bar is proportional to the contribution of the corresponding site. Sites above the threshold located inside the coding regions are shown in red. Site positions in each protein are summarized in the table. (b) Changes in frequencies of the sites that contribute most, which better reflect the geographical structure in each subpopulation (pop1 to pop6).

or positively charged arginine (alleles G or T), with arginine being predominant in pop3 and pop4, an overlap of both amino acids in pop5, and glycine in pop1, pop2 and pop6 (Figs 4b and 5).

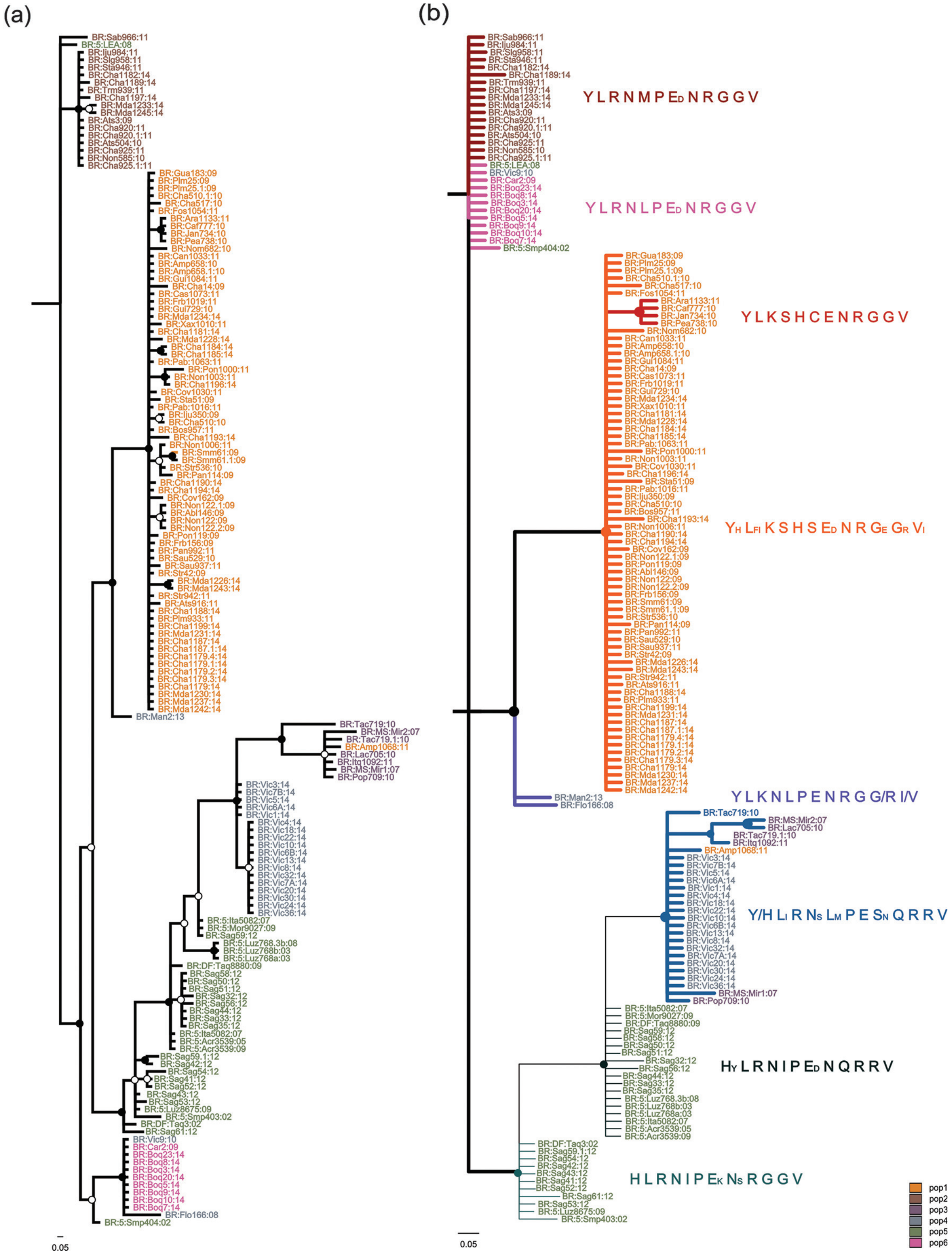


Fig. 5. Bayesian inference tree based on the sites of EuYMV DNA-A that contribute most to geographical structure. Nodes to the right of branches with posterior probabilities equal to or higher than 0.8 are indicated by filled circles and those with values lower than 0.8 and higher 0.5 are denoted by empty circles. Isolate colours based on DAPC-inferred subpopulations. Scale bars indicate substitutions per site. (a) Tree based on codon sequences of the sites that contribute most. (b) Tree based on variable sites of amino acid sequences of the sites that contribute most, with branches coloured according to amino acid sequences. Amino acids correspond to (amino acid site/protein): 1, 223/*CP*; 2, 80/*Rep*; 3, 149/*Rep*; 4, 160/*Rep*; 5, 172/*Rep*; 6, 180/*Rep*; 7, 255/*Rep*; 8, 16/*Ren*; 9, 78/*Trap*; 10, 33/*Ren*; 11, 80/*Trap*; 12, 100/*Ren*. Small letters correspond to amino acids that occur in less than three sequences, and forward slashes correspond to amino acids which occur with similar frequencies.

The polymorphisms lead to non-synonymous substitutions in the remaining 11 informative sites (Table S8). Polymorphic sites 1056, 2007, 2037 and 2244 displayed non-synonymous substitutions, but involved exchanges between amino acids with the same biochemical properties. For example, codon 149 (site 2037 in *Rep*) encoded positively charged amino acids, with a predominance of arginine (allele T) in all subpopulations except pop1, where lysine (allele C) was predominant (Figs 4b, 5, and Table S8). Further, strong evidence for negative selection in codon 149 was found in pop1 and pop5, and weak evidence for positive selection was found in pop4 (where the frequency of allele C was >1).

However, the substitutions in sites 1259, 1718, 1943, 1945, 1967, 1968 and 1969 lead to exchanges between different amino acid types (Fig. 4b and Table S8). For example, polymorphisms in sites 1967, 1968 and 1969 corresponding to codon 172 in *Rep* yielded translation of either hydrophilic non-charged (methionine or histidine) or hydrophobic (leucine or isoleucine) amino acids. Histidine, methionine and isoleucine were prevalent in pop1, pop2 and pop5, respectively. A balance between the frequencies of methionine and histidine (five isolates encoding leucine and two encoding methionine) was found in pop3, while leucine was predominant in pop4 and pop6 (Figs 4b, 5). Translation of amino acids from different classes in different subpopulations suggests that different protein features are being selected at each sampling location, contributing to the genetic divergence among isolates from different subpopulations.

DISCUSSION

The introduction of *B. tabaci* MEAM1 in Brazil has facilitated the transfer of indigenous begomoviruses from non-cultivated plants to economically important crops, something that has been surveyed extensively [11–13, 25, 26, 29, 34, 35]. While most studies have focused on viral species diversity, others have investigated the genetic structure of begomovirus populations infecting both cultivated and non-cultivated hosts in different geographical regions of the country [12, 13, 29, 31, 36]. Nevertheless, very few studies have investigated the genetic variability of viruses that remain restricted to non-cultivated hosts. Knowledge of the population dynamics of these viruses is important, since they may act as reservoirs of virus diversity, where intraspecific recombination and pseudorecombination may occur [13, 26, 29].

Since the first report of EuYMV infecting *E. heterophylla* in Brazil [24], the virus has been described in other non-cultivated plants [12, 25, 26]. So far, only Barreto *et al.* [26] found a begomovirus other than EuYMV infecting *E. heterophylla* plants (ToSRV), and even then in a mixed infection with EuYMV and accumulating at a low titer. Likewise, in our study, which comprised an extensive survey in sampling sites covering states from the south to the north of Brazil, EuYMV was the only begomovirus found.

Previous studies found a high level of genetic variability in begomovirus populations infecting non-cultivated hosts. Compared with *Macroptilium yellow spot virus* (MaYSV; π=0.06580 for the full-length of DNA-A), EuYMV has a lower degree of variability [31, 36]. We found that the EuYMV DNA-B showed higher genetic variability compared with the DNA-A, in agreement with previous studies, which suggested that bipartite begomovirus components have distinct evolutionary histories [30].

A number of previous studies have indicated that EuYMV belongs to a phylogenetic lineage distinct from other begomoviruses isolated in Brazil. Silva *et al.* [12] isolated EuYMV from *Macroptilium atropurpureum*, and found EuYMV clustering with viruses from Central America. Tavares *et al.* [25] isolated EuYMV from *Sida santaremnensis*, and indicated that EuYMV is related to a group comprised mostly by begomovirus found in Central and North America. Similarly, Fernandes *et al.* [24] isolated EuYMV from *E. heterophylla*, and described EuYMV isolates as more closely related to Peruvian and North American begomoviruses than to Brazilian begomoviruses, hypothesizing that EuYMV was introduced into Brazil. These authors placed EuYMV in the so-called *Squash leaf curl virus* (SqLCV) clade [24]. Some species of this clade have distinct but overlapping host ranges, and Idris *et al.* [37] demonstrated that members of the SqLCV clade can form viable pseudorecombinants able to extend their host ranges. Rocha *et al.* [29] indicated the clustering of EuYMV with Central and North American begomovirus, and also pointed to a recombinant origin of EuYMV, detecting an interspecific recombination event with a major parent from Central America. In our DNA-A dataset, which only comprises EuYMV sequences, no recombination events were detected. Thus, although recombination seems to be involved in the origin of EuYMV, it appears that no intraspecies recombination occurred in the DNA-A once the viral population established itself in Brazil. However, it must be pointed out that the low degree of genetic variability amongst EuYMV isolates makes it difficult to distinguish between recombinants and parental sequences.

In contrast to the DNA-A dataset, five recombinant isolates were detected in the DNA-B dataset, with two events involving breakpoints located within coding regions, which are considered to be ‘cold spots’ for recombination among begomoviruses [38]. These recombination events could explain the higher variability of the DNA-B compared with the DNA-A.

Information on the intraspecific genetic variability of begomoviruses is relevant to provide clues about virulence and dispersion [12, 29, 32, 36, 39–43]. We assessed the genetic variability of EuYMV, investigated the factors that determine the genetic structure and described how the genetic variability is distributed within the subpopulations. The distribution of variability, generated by mutation or recombination, in the viral populations depends on genetic drift and selection, but it is often difficult to differentiate between the effects of genetic drift and those of selection [40]. Because EuYMV is essentially the only virus found in *E. heterophylla* and is only rarely found in other hosts, it is a good viral system to study the effects of genetic drift and selection in the population structure, since it experiences less influence from adaptive selection to different hosts.

We tried to highlight how these evolutionary processes could shape the genetic variability of EuYMV populations, using the information on variability between subpopulations to identify the genome sites that contribute most to the geographical structure proposed, and studying the effects of selection while considering the contribution of polymorphic sites within coding regions.

A great effort has been made to define begomovirus population differentiation according to geographical origin. Some studies, using Bayesian clustering approaches implemented in the program STRUCTURE, aimed to identify the population structure at a global [32] or local [29] scale. We provided the first analysis comparing the population structure in a large sampled area using two different methods: Bayesian clustering and multivariate statistical analysis. For simulated data, multivariate analysis using DAPC proved to be as accurate as STRUCTURE [44]. Using STRUCTURE, we were able to differentiate two subpopulations based on sampling locations, which corresponded to the two major clades inferred by phylogenetic analysis. However, the application of DAPC seemed to be more suitable for our dataset, since it also allowed the differentiation of six subpopulations according to sampling locations and in agreement with phylogenetic analysis.

Similar to begomoviruses, the cryptic species complex of *B. tabaci* exhibits a strong geographical pattern in a global context [45]. However, studies analysing the population structure within each species of *B. tabaci* at a local scale failed to find evidence of isolation by distance. Such non-geographical structuring could indicate long-distance migrations [46, 47]. De Barro [48] detected six genetic populations in the Asia–Pacific region with little or no gene flow between them. It is possible that the migration is limited by the year-round availability of hosts in tropical regions [47]. Low genetic diversity was observed in NW populations in comparison with OW populations [45]. In Brazil, *B. tabaci* MEAM1 is prevalent, but the New World 1 (NW1) and New World 2 (NW2) species have also been reported in different regions [49], and the Mediterranean (MED) species was recently reported in the southern region [50]. A lack of information about the genetic structure of *B. tabaci* populations in Brazil prevented us from comparing the viral population structure in relation to (or as a function of) that of *B. tabaci*.

Genetic variability was higher in subpopulations from regions where the agricultural landscape has a greater diversity of hosts, for example pop3 from MS. Environmental heterogeneity could also modulate the genetic structure of both host and virus. Landscape heterogeneity affecting the genetic structure of viral populations was demonstrated for two bipartite begomoviruses, *Pepper golden mosaic* (PepGMV) and *Pepper huasteco yellow vein virus* (PHYVV), by comparing neighbouring populations of wild and human-managed chiltepin (*Capsicum annuum glabriusculum*), a host with a large degree of genetic variability and strong spatial structure [51]. The prevalence of each virus was significantly higher in cultivated than in wild populations. Nevertheless, the effect of ecosystem biodiversity on the genetic diversity depended on the virus species. The loss of biodiversity at higher levels of habitat anthropization was associated with increased genetic diversity of PepGMV but not of PHYVV [52]. A possible influence of *E. heterophylla* diversity and agro-climate environment in the EuYMV population structure will be addressed in our future studies.

DAPC analysis allowed the visual assessment of between-population differentiation and the contribution of individual alleles to population structuring. The genome sites that were most markedly different across the proposed subpopulations were highlighted as contributing the most to genetic divergence. These sites were mapped in the genome and compared at the level of amino acid changes. We found each gene/subpopulation to be under different selective pressures, albeit with a tendency for purifying selection to act upon each subpopulation. Many more sites were evolving under negative selection compared to positive selection, in agreement with previous studies that demonstrated negative selection predominating in coding regions during the evolution of plant viruses [29, 31, 39, 40]. In general, negative selection was predominant in all six subpopulations.

We were able to reconstruct the phylogenetic analysis solely using the sites that contributed most, demonstrating that these polymorphisms hold supporting information to discriminate between subpopulations and may contain valuable clues to understand the main evolutionary mechanisms underlying genetic differentiation. Considering the predominance of negative selection acting on the EuYMV populations, the polymorphisms were ranked according to three descriptions: synonymous substitutions, non-synonymous substitutions with the same type of amino acid and non-synonymous substitutions with a different type of amino acid. The mechanism underlying substitutions was most likely genetic drift, with different alleles being fixed or changing in frequency, yielding random fluctuations in the number of variants present in each subpopulation. On the other hand, different types of amino acids being translated in each subpopulation could be evidence of distinct features being selected at each sampling location, contributing to the genetic divergence among isolates from different subpopulations. The importance of these polymorphisms in the adaptability of each isolate to different environmental conditions could be assessed by directional mutation in infectious clones.

In conclusion, EuYMV has a lower degree of genetic variability compared with other begomovirus populations infecting non-cultivated plants, and only a few intraspecific recombination events (restricted to the DNA-B dataset) were detected. EuYMV displays a different pattern from that commonly observed in begomovirus populations (high variability and frequent recombination events), but nevertheless, similar to other begomoviruses, segregates according to sampling location. It could be a good system to study the standing genetic variability of begomovirus populations and how genetic drift and selection contribute to maintain the patterns of begomovirus population diversity from a geographical structuring perspective.

METHODS

Sampling

Samples of *E. heterophylla* plants showing typical symptoms of begomovirus infection (yellow mosaic, leaf curling and stunting) were collected in locations throughout the states of Amazonas (AM), Goiás (GO), Mato Grosso do Sul (MS), Minas Gerais (MG), Paraíba (PB), Pernambuco (PE), Paraná (PR), Rio Grande do Sul (RS) and Santa Catarina (SC) and in the Federal District (DF) from 2009 to 2014 (Tables S1 and S2). Samples from southern Brazil (RS, PR, SC and MS) were mainly collected within and near maize, soybean and wheat fields. The MS agricultural landscape has a greater diversity of hosts. Samples from GO/DF and from the northeastern states (PB and PE) were collected within and near common bean fields. Samples from AM and MG were collected in natural wild habitats. For each sample the following information was recorded: date of collection, GPS coordinates of the sampling location and symptoms (description and digital image of the sample at the time of collection). The samples were stored in plastic bags and transported to the laboratory where they were press-mounted until DNA extraction [53].

Cloning and sequencing of full-length begomovirus genomes

Viral genomes were amplified using rolling-circle amplification (RCA) according to Inoue-Nagata *et al.* [54]. Aliquots of the amplification products were subjected to cleavage with restriction enzymes to obtain fragments of approximately 2600 nucleotides (nt), corresponding to one genomic copy of each DNA component. These fragments were cloned into the pBLUESCRIPT-KS+ (Stratagene) plasmid vector and completely sequenced at Macrogen, Inc. (Seoul, Republic of Korea). Species assignment was based on the ICTV-established cut-off of 91 % nt sequence identity for the full-length DNA-A [55] using pairwise comparisons performed with SDT v. 1.2 [56].

Recombination and phylogenetic analysis

Multiple sequence alignments were performed using the MUSCLE algorithm [57]. Partitions in viral genomes with conflicting evolutionary histories were detected using RDP, Geneconv, Bootscan, Maximum Chi Square, Chimaera, SisterScan and 3Seq methods, as implemented in the RDP4 program [58], using default settings for each analysis method and a Bonferroni-corrected *P*-value of 0.05. Only recombination events detected by at least four methods were considered reliable. Recombinant blocks were eliminated from the data sets. Bayesian analysis was performed with MrBayes 3.0 [59] using the models selected by MrModeltest2.2 [60] in the Akaike information criterion (AIC). Two independent analyses were conducted, each running at least 10 000 000 generations. Phylogenetic trees were visualized and edited using FigTree 1.3 (http://tree.bio.ed.ac.uk/software/figtree/).

Genetic structure of the viral population

The main descriptors of molecular variability were estimated using DnaSP v. 5 [61]. The average pairwise number of nucleotide differences per site (nucleotide diversity, π) was estimated using a sliding window of 100 nt, with a step size of 10 nt. The statistical significance of the differences amongst the mean nucleotide diversity obtained from different datasets was calculated according to Lima *et al.* [62]. The correlation between distance matrices was tested by Mantel's test using the vegan package in R software [63]. Two methods were used to infer population subdivision and assign individuals to each subpopulation. To support the assumption of linkage equilibrium for STRUCTURE analysis, a null hypothesis of linkage equilibrium was tested by Monte Carlo simulations using the program LIAN v. 3.7 [64] as proposed by Prasanna *et al.* [32]. Bayesian clustering using STRUCTURE [65] was performed in ten independent runs of 1 000 000 Markov chain Monte Carlo (MCMC) replications with a burn-in of 10 000 runs for each *K*-value varying from 1 to 10. The suitable *K*-values were determined by the higher likelihood of [Ln P (D)] [66]. For a better visualization of the genetic structure, the ancestry coefficients were plotted in a map generated in Philcarto software [67]. Multivariate statistical analysis using discriminant analysis of principal components (DAPC) was performed with the package adegenet implemented in R software [44]. Preliminary subpopulations were inferred by *K*-means run sequentially with increasing values of *K*, retaining all principal components (PCs), and different clustering solutions were compared using the Bayesian information criterion (BIC). After assigning individuals to each inferred subpopulation, the optimum PC values were investigated to assess the percentage of successful reassignment, a-scores and cross-validation. Wright's F fixation index [68] was estimated for the distinct inferred subpopulations with 10 000 replications in DnaSP v. 5 [61] to support each inferred subpopulation. In addition, different population subdivision hypotheses were tested by analysis of molecular variance (AMOVA) implement in Arlequin v. 3.5 [69] and the proper assignment of each isolate into a subpopulation was checked by spatial analysis of molecular variance (SAMOVA), defining populations that are geographically homogeneous and maximally differentiated from each other [70]. To provide insights into the underlying causes of maintenance of begomovirus geographical structuring we inspected the associated site loadings. As proposed by Jombart [44], the contributions of sites were plotted for a single DAPC discriminant function at a time. We scattered the density of individuals in each subpopulation, and computed the contribution of sites given a threshold of 0.01. Assuming only sites covering coding regions, the frequencies of the 23 most contributing sites were plotted according to geographical structuring, mapped in the genome and compared at the level of amino acid changes. We also reconstructed a Bayesian-inferred tree based on variable amino acids of the sites that contributed most.

Detection of negative and positive selection

Three types of neutrality tests were used to test for the occurrence of selection in populations: Tajima's D and Fu and Li's D* and F*. The analyses were performed with DnaSP v. 5 [61]. Amino acid sites under selection were investigated with the maximum likelihood based-methods available in the DataMonkey webserver [71], after determining nt substitution models in MODELTEST and screening for recombination using GARD [72]. The SLAC method was used to estimate the mean ratios of non-synonymous to synonymous substitutions (d_N/d_S) for each open reading frame in the DNA-A components of each EuYMV subpopulation.

Funding information
This work was funded by CNPq (grant 401838/2013-7 to F. M. Z.) and FAPEMIG (grant CAG-APQ-02037-13 to F. M. Z.) and by Embrapa (grant SEG 02.08.01.008.00.00). T. B. M. was the recipient of a CNPq doctoral fellowship.

Acknowledgements
The authors wish to thank Jesus Navas-Castillo, Elvira Fiallo-Olivé and Eduardo S.G. Mizubuti for helpful discussions and critical reading of the manuscript.

Conflicts of interest
The authors declare that there are no conflicts of interest.

Ethical statement
The research reported in this paper did not involve animals or humans.

References

1. Zerbini FM, Briddon RW, Idris A, Martin DP, Moriones E *et al.* ICTV Virus Taxonomy Profile: *Geminiviridae*. *J Gen Virol* 2017;98:131–133.
2. Harrison BD, Robinson DJ. Natural genomic and antigenic variation in white-fly transmitted geminiviruses (begomoviruses). *Annu Rev Phytopathol* 1999;39:369–398.
3. Rojas MR, Hagen C, Lucas WJ, Gilbertson RL. Exploiting chinks in the plant's armor: evolution and emergence of geminiviruses. *Annu Rev Phytopathol* 2005;43:361–394.
4. Duffy S, Holmes EC. Phylogenetic evidence for rapid rates of molecular evolution in the single-stranded DNA begomovirus *Tomato yellow leaf curl virus*. *J Virol* 2008;82:957–965.
5. Duffy S, Holmes EC. Validation of high rates of nucleotide substitution in geminiviruses: phylogenetic evidence from East African cassava mosaic viruses. *J Gen Virol* 2009;90:1539–1547.
6. Padidam M, Sawyer S, Fauquet CM. Possible emergence of new geminiviruses by frequent recombination. *Virology* 1999;265:218–225.
7. Andrade EC, Manhani GG, Alfenas PF, Calegario RF, Fontes EP *et al.* *Tomato yellow spot virus*, a tomato-infecting begomovirus from Brazil with a closer relationship to viruses from *Sida* sp., forms pseudorecombinants with begomoviruses from tomato but not from *Sida*. *J Gen Virol* 2006;87:3687–3696.
8. Jones DR. Plant viruses transmitted by whiteflies. *Eur J Plant Pathol* 2003;109:195–219.
9. Dinsdale A, Cook L, Riginos C, Buckley YM, de Barro P. Refined global analysis of *Bemisia tabaci* (Hemiptera: Sternorrhyncha: Aleyrodoidea: Aleyrodidae) mitochondrial cytochrome oxidase 1 to identify species level genetic boundaries. *Ann Entomol Soc Am* 2010;103:196–208.
10. Navas-Castillo J, Fiallo-Olivé E, Sánchez-Campos S. Emerging virus diseases transmitted by whiteflies. *Annu Rev Phytopathol* 2011;49:219–248.
11. Castillo-Urquiza GP, Beserra JE Jr, Bruckner FP, Lima AT, Varsani A *et al.* Six novel begomoviruses infecting tomato and associated weeds in Southeastern Brazil. *Arch Virol* 2008;153: 1985–1989.
12. Silva SJC, Castillo-Urquiza GP, Hora-Júnior BT, Assunção IP, Lima GSA *et al.* Species diversity, phylogeny and genetic variability of begomovirus populations infecting leguminous weeds in northeastern Brazil. *Plant Pathol* 2012;61:457–467.
13. Silva FN, Lima AT, Rocha CS, Castillo-Urquiza GP, Alves-Júnior M *et al.* Recombination and pseudorecombination driving the evolution of the begomoviruses *Tomato severe rugose virus* (ToSRV) and *Tomato rugose mosaic virus* (ToRMV): two recombinant DNA-A components sharing the same DNA-B. *Virol J* 2014;11:66.
14. Pinto VB, Silva JP, Fiallo-Olivé E, Navas-Castillo J, Zerbini FM. Novel begomoviruses recovered from *Pavonia* sp. in Brazil. *Arch Virol* 2016;161:735–739.
15. Fiallo-Olivé E, Zerbini FM, Navas-Castillo J. Complete nucleotide sequences of two new begomoviruses infecting the wild malvaceous plant *Melochia* sp. in Brazil. *Arch Virol* 2015;160:3161–3164.
16. Awadalla P. The evolutionary genomics of pathogen recombination. *Nat Rev Genet* 2003;4:50–60.
17. Sattar MN, Kvarnheden A, Saeed M, Briddon RW. Cotton leaf curl disease – an emerging threat to cotton production worldwide. *J Gen Virol* 2013;94:695–710.
18. Lefeuvre P, Moriones E. Recombination as a motor of host switches and virus emergence: geminiviruses as case studies. *Curr Opin Virol* 2015;10:14–19.
19. Wilson AK. *Euphorbia heterophylla:* a review of distribution, importance and control. *Trop Pest Manag* 1981;27:32–38.
20. Cronquist A. *An Integrated System of Classification of Flowering Plants*. New York: Columbia University Press; 1981. p. 1262.
21. Vidal RA, Winkler LM. *Euphorbia heterophylla* L. resistant to herbicide inhibitors of acetolactate synthase: II - Geographic distribution and genetic characterization of biotypes from Rio Grande do Sul plains. *Rev Bras Agroc* 2004;10:461–465.
22. Christoffoleti PJ. *Aspectos Da Resistência De Plantas Daninhas a Herbicidas*, 3rd ed. Piracicaba: HRAC-BR; 2008. p. 120.
23. Costa AS, Bennett CW. Whitefly transmitted mosaic of *Euphorbia prunifolia*. *Phytopathology* 1950;40:266–283.
24. Fernandes FR, Albuquerque LC, de Oliveira CL, Cruz AR, da Rocha WB *et al.* Molecular and biological characterization of a new Brazilian begomovirus, euphorbia yellow mosaic virus (EuYMV), infecting *Euphorbia heterophylla* plants. *Arch Virol* 2011; 156:2063–2069.
25. Tavares SS, Ramos-Sobrinho R, González-Aguilera J, Lima GSA, Assunção IP *et al.* Further molecular characterization of weed-associated begomoviruses in Brazil with an emphasis on *Sida* spp. *Planta Daninha* 2012;30:305–315.
26. Barreto SS, Hallwass M, Aquino OM, Inoue-Nagata AK. A study of weeds as potential inoculum sources for a tomato-infecting begomovirus in central Brazil. *Phytopathology* 2013;103:436–444.
27. Richter KS, Ende L, Jeske H. Rad_{54} is not essential for any geminiviral replication mode *in planta*. *Plant Mol Biol* 2015;87:193–202.
28. Sottoriva LD, Lourenção AL, Colombo CA. Performance of *Bemisia tabaci* (Genn.) biotype B (Hemiptera: Aleyrodidae) on weeds. *Neotrop Entomol* 2014;43:574–581.
29. Rocha CS, Castillo-Urquiza GP, Lima AT, Silva FN, Xavier CA *et al.* Brazilian begomovirus populations are highly recombinant, rapidly evolving, and segregated based on geographical location. *J Virol* 2013;87:5784–5799.
30. Briddon RW, Patil BL, Bagewadi B, Nawaz-Ul-Rehman MS, Fauquet CM. Distinct evolutionary histories of the DNA-A and DNA-B components of bipartite begomoviruses. *BMC Evol Biol* 2010;10:97.
31. Lima AT, Sobrinho RR, González-Aguilera J, Rocha CS, Silva SJ *et al.* Synonymous site variation due to recombination explains

higher genetic variability in begomovirus populations infecting non-cultivated hosts. *J Gen Virol* 2013;94:418–431.

32. **Prasanna HC, Sinha DP, Verma A, Singh M, Singh B *et al.*** The population genomics of begomoviruses: global scale population structure and gene flow. *Virol J* 2010;7:220.

33. **Jombart T, Devillard S, Balloux F.** Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet* 2010;11:94.

34. **Albuquerque LC, Varsani A, Fernandes FR, Pinheiro B, Martin DP *et al.*** Further characterization of tomato-infecting begomoviruses in Brazil. *Arch Virol* 2012;157:747–752.

35. **Fernandes FR, Cruz AR, Faria JC, Zerbini FM, Aragão FJ.** Three distinct begomoviruses associated with soybean in central Brazil. *Arch Virol* 2009;154:1567–1570.

36. **Sobrinho RR, Xavier CA, Pereira HM, Lima GS, Assunção IP *et al.*** Contrasting genetic structure between two begomoviruses infecting the same leguminous hosts. *J Gen Virol* 2014;95:2540–2552.

37. **Idris AM, Mills-Lujan K, Martin K, Brown JK.** *Melon chlorotic leaf curl virus*: characterization and differential reassortment with closest relatives reveal adaptive virulence in the *squash leaf curl virus* clade and host shifting by the host-restricted *bean calico mosaic virus*. *J Virol* 2008;82:1959–1967.

38. **Lefeuvre P, Lett JM, Varsani A, Martin DP.** Widely conserved recombination patterns among single-stranded DNA viruses. *J Virol* 2009;83:2697–2707.

39. **Nouri S, Arevalo R, Falk BW, Groves RL.** Genetic structure and molecular variability of *Cucumber mosaic virus* isolates in the United States. *PLoS One* 2014;9:e96582.

40. **García-Arenal F, Fraile A, Malpica JM.** Variability and genetic structure of plant virus populations. *Annu Rev Phytopathol* 2001; 39:157–186.

41. **González-Aguilera J, Tavares SS, Sobrinho RR, Xavier CAD, Dueñas-Hurtado F *et al.*** Genetic structure of a brazilian population of the begomovirus *Tomato severe rugose virus* (ToSRV). *Tropical Plant Pathology* 2012;37:346–353.

42. **Yang XL, Zhou MN, Qian YJ, Xie Y, Zhou XP.** Molecular variability and evolution of a natural population of tomato yellow leaf curl virus in Shanghai, China. *J Zhejiang Univ Sci B* 2014;15:133–142.

43. **Acosta-Leal R, Duffy S, Xiong Z, Hammond RW, Elena SF.** Advances in plant virus evolution: translating evolutionary insights into better disease management. *Phytopathology* 2011;101:1136–1148.

44. **Jombart T.** *adegenet*: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 2008;24:1403–1405.

45. **Hadjistylli M, Roderick GK, Brown JK.** Global population structure of a worldwide pest and virus vector: genetic diversity and population history of the *Bemisia tabaci* sibling species group. *PLoS One* 2016;11:e0165105.

46. **Dalmon A, Halkett F, Granier M, Delatte H, Peterschmitt M.** Genetic structure of the invasive pest *Bemisia tabaci*: evidence of limited but persistent genetic differentiation in glasshouse populations. *Heredity (Edinb)* 2008;100:316–325.

47. **Tahiri A, Halkett F, Granier M, Gueguen G, Peterschmitt M.** Evidence of gene flow between sympatric populations of the Middle East-Asia Minor 1 and Mediterranean putative species of *Bemisia tabaci*. *Ecol Evol* 2013;3:2619–2633.

48. **De Barro PJ.** Genetic structure of the whitefly *Bemisia tabaci* in the Asia–Pacific region revealed using microsatellite markers. *Mol Ecol* 2005;14:3695–3718.

49. **Marubayashi JM, Yuki VA, Rocha KCG, Mituti T, Pelegrinotti FM *et al.*** At least two indigenous species of the *Bemisia tabaci* complex are present in Brazil. *J Appl Entomol* 2013;137:113–121.

50. **da Fonseca Barbosa L, Yuki VA, Marubayashi JM, de Marchi BR, Perini FL *et al.*** First report of *Bemisia tabaci* Mediterranean (Q biotype) species in Brazil. *Pest Manag Sci* 2015;71:501–504.

51. **Rodelo-Urrego M, Pagán I, González-Jara P, Betancourt M, Moreno-Letelier A *et al.*** Landscape heterogeneity shapes host-parasite interactions and results in apparent plant-virus codivergence. *Mol Ecol* 2013;22:2325–2340.

52. **Rodelo-Urrego M, García-Arenal F, Pagán I.** The effect of ecosystem biodiversity on virus genetic diversity depends on virus species: a study of chiltepin-infecting begomoviruses in Mexico. *Virus Evol* 2015;1:vev004.

53. **Doyle JJ, Doyle JL.** A rapid DNA isolation procedure for small amounts of fresh leaf tissue. *Phytochem Bull* 1987;19:11–15.

54. **Inoue-Nagata AK, Albuquerque LC, Rocha WB, Nagata T.** A simple method for cloning the complete begomovirus genome using the bacteriophage phi$_{29}$ DNA polymerase. *J Virol Methods* 2004;116:209–211.

55. **Brown JK, Zerbini FM, Navas-Castillo J, Moriones E, Ramos-Sobrinho R *et al.*** Revision of *Begomovirus* taxonomy based on pairwise sequence comparisons. *Arch Virol* 2015;160:1593–1619.

56. **Muhire BM, Varsani A, Martin DP.** SDT: a virus classification tool based on pairwise sequence alignment and identity calculation. *PLoS One* 2014;9:e108277.

57. **Edgar RC.** MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 2004;5:113.

58. **Martin DP, Murrell B, Golden M, Khoosal A, Muhire B.** RDP4: detection and analysis of recombination patterns in virus genomes. *Virus Evol* 2015;1:vev003.

59. **Ronquist F, Huelsenbeck JP.** MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 2003;19:1572–1574.

60. **Nylander JAA** MrModeltest v2. *Program distributed by the author Evolutionary Biology Centre*, Uppsala University; 2004

61. **Rozas J, Sánchez-Delbarrio JC, Messeguer X, Rozas R.** DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 2003;19:2496–2497.

62. **Lima ATM, Silva JCF, Silva FN, Castillo-Urquiza GP, Silva FF *et al.*** The diversification of begomovirus populations is predominantly driven by mutational dynamics. *Virus Evol* 2017;3:vex005.

63. **Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR *et al.*** 2013. Vegan: community Ecology Package. R package version 2.0-7. http://CRANR-projectorg/package=vegan [Accessed on May 22, 2017].

64. **Haubold B, Hudson RR.** LIAN 3.0: detecting linkage disequilibrium in multilocus data. Linkage analysis. *Bioinformatics* 2000;16:847–849.

65. **Hubisz MJ, Falush D, Stephens M, Pritchard JK.** Inferring weak population structure with the assistance of sample group information. *Mol Ecol Resour* 2009;9:1322–1332.

66. **Earl DA, Vonholdt BM.** STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the evanno method. *Conserv Genet Resour* 2012;4:359–361.

67. **Waniez P.** Philcarto: histoire de vie d'un logiciel de cartographie. *Eur J Geo* 2010:497.

68. **Weir BS.** *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*. Sunderland, Massachusetts: Sinauer Associated Inc; 1996. p. 445.

69. **Excoffier L, Lischer HE.** Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour* 2010;10:564–567.

70. **Dupanloup I, Schneider S, Excoffier L.** A simulated annealing approach to define the genetic structure of populations. *Mol Ecol* 2002;11:2571–2581.

71. **Delport W, Poon AF, Frost SD, Kosakovsky Pond SL.** Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* 2010;26:2455–2457.

72. **Martin DP, Posada D, Crandall KA, Williamson C.** A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. *AIDS Res Hum Retroviruses* 2005;21:98–102.