Synonymous site variation due to recombination explains higher genetic variability in begomovirus populations infecting non-cultivated hosts

Alison T. M. Lima,1 Roberto R. Sobrinho,1 Jorge González-Aguilera,1 Carolina S. Rocha,1‡ Sarah J. C. Silva,1§ César A. D. Xavier,1 Fábio N. Silva,1 Siobain Duffy2 and F. Murilo Zerbini1

Correspondence
Siobain Duffy
duffy@aesop.rutgers.edu
Francisco Murilo Zerbini
zerbini@ufv.br

Departamento de Fitopatologia/BIOAGRO, Universidade Federal de Viçosa, Viçosa, MG 36570-000, Brazil

Department of Ecology, Evolution and Natural Resources, Rutgers, The State University of New Jersey, New Brunswick, NJ 08901, USA

Begomoviruses are ssDNA plant viruses that cause serious epidemics in economically important crops worldwide. Non-cultivated plants also harbour many begomoviruses, and it is believed that these hosts may act as reservoirs and as mixing vessels where recombination may occur. Begomoviruses are notoriously recombination-prone, and also display nucleotide substitution rates equivalent to those of RNA viruses. In Brazil, several indigenous begomoviruses have been described infecting tomatoes following the introduction of a novel biotype of the whitefly vector in the mid-1990s. More recently, a number of viruses from non-cultivated hosts have also been described. Previous work has suggested that viruses infecting non-cultivated hosts have a higher degree of genetic variability compared with crop-infecting viruses. We intensively sampled cultivated and non-cultivated plants in similarly sized geographical areas known to harbour either the weed-infecting Macroptilium yellow spot virus (MaYSV) or the crop-infecting Tomato severe rugose virus (ToSRV), and compared the molecular evolution and population genetics of these two distantly related begomoviruses. The results reinforce the assertion that infection of non-cultivated plant species leads to higher levels of standing genetic variability, and indicate that recombination, not adaptive selection, explains the higher begomovirus variability in non-cultivated hosts.

INTRODUCTION

Single-stranded DNA begomoviruses (whitefly transmitted members of the family Geminiviridae) have become an important factor limiting crop production in tropical and subtropical regions (Morales & Anderson, 2001; Rojas et al., 2005; Seal et al., 2006). Begomoviruses causing cassava mosaic disease (CMD) are the major biotic constraint to cassava cultivation in Africa (Legg & Fauquet, 2004; Legg & Thresh, 2000; Ndagugu et al., 2005; Were et al., 2004). A complex of at least six different begomovirus species is responsible for the devastating tomato yellow leaf curl disease (TYLCD) (Moriones & Navas-Castillo, 2000). In the Americas, diseases caused by begomoviruses have significantly impacted tomato and bean production since the 1980s (Blair et al., 1995; Brown & Bird, 1992; Gilbertson et al., 1993; Morales & Jones, 2004; Polston & Anderson, 1997).

Previous studies have indicated a high intra- and interspecific diversity of begomoviruses, which can facilitate adaptation to new climates and novel hosts (Monci et al., 2002). Recombination is the most intensively studied population genetic process in begomoviruses, and has been considered more significant than mutation by many researchers (Lefeuvre et al., 2007a, 2009; Martin et al., 2005, 2011; Monci et al., 2002; Padidam et al., 1999; Pita et al., 2001). It appears to heavily contribute to begomovirus genetic diversity, increasing the evolutionary potential and local adaptation of strains (Berrie et al., 2001; Graham et al., 2010; Harrison & Robinson, 1999; Monci

1Present address: Embrapa Trigo, Rodovia BR 285 km 294, CP 451, Passo Fundo, RS, 99001-970, Brazil.
2Present address: FuturaGene Brasil, Avenida José Lembo 1010, Itapetinga, SP, 18210-780, Brazil.
3Present address: Dep. de Fitossanidade/CECA, Universidade Federal de Alagoas, Rio Largo, AL, 57100-000, Brazil.

The GenBank/EMBL/DBJ accession numbers for the sequences reported in this paper are KC004091–KC004134, KC004068–KC004090 and JX865615–JX865650.

Supplementary material is available with the online version of this paper.
et al., 2002; Padidam et al., 1999). There is ample opportunity for recombination because multiple begomovirus species are often found co-infecting the same plant (Davino et al., 2009; García-Andrés et al., 2006; Harrison et al., 1997; Pita et al., 2001; Ribeiro et al., 2003; Sanz et al., 2000; Torres-Pacheco et al., 1996), and more than one virus can simultaneously replicate in the same nucleus (Morilla et al., 2004). Additionally, the high recombination frequency observed for begomoviruses may be explained by a theoretical recombination-dependent replication mechanism (RDR) (Jeske et al., 2001), in addition to the well-documented rolling circle replication (RCR) (Saunders et al., 2001). However, recent studies have also indicated that begomoviruses can evolve by mutation alone as quickly as RNA viruses (Duffy & Holmes, 2008, 2009), and positive selection – on mutations or the products of recombination events – may also play a role in begomovirus evolutionary dynamics (Monci et al., 2002; Pita et al., 2001; Zhou et al., 1997).

Additionally, host use may play an important role in the standing genetic variability of begomovirus populations (Seal et al., 2006). Several species of non-cultivated plants, especially of the families Malvaceae, Euphorbiaceae, Fabaceae and Solanaceae, are known hosts of begomoviruses (Morales & Anderson, 2001). These weed/wild hosts can serve as reservoirs for infection of nearby crops (Alabi et al., 2008; Barbosa et al., 2009; Bedford et al., 1998; García-Andrés et al., 2006), as overwintering refugia (Alabi et al., 2007, 2008; García-Andrés et al., 2006) and as ‘mixing vessels’ for interspecific coinfection and recombination (García-Andrés et al., 2006; Monde et al., 2010; Silva et al., 2012). Increased host use and diminished bottlenecks would both potentially increase the effective population size of begomovirus populations (Power, 2000; Seal et al., 2006). Although there is limited data on the variability of begomovirus populations in non-cultivated hosts, such data suggest that it is higher than that observed in crop-infecting begomoviruses (Fiallo-Olivé et al., 2012; Silva et al., 2011, 2012; Wyant et al., 2011).

To gather data on the factors affecting genetic variability in begomovirus populations and shed light on whether frequent infection of non-cultivated plants alters viral evolutionary dynamics, we contrasted two populations of distantly related begomoviruses. We intensively sampled crops and non-cultivated plants in similarly sized geographical areas known to harbour either Macropitilium lathyroides (MaYSV) or Tomato severe rugose virus (ToSRV). MaYSV is a recently isolated species that was previously only reported in non-cultivated hosts in north-eastern Brazil (Silva et al., 2012), whereas ToSRV is the most widespread tomato-infecting begomovirus in the country (Fernandes et al., 2008; Rocha, 2011; Zerbini et al., 2005). ToSRV can naturally infect other important crops such as chili pepper (Bezerra-Agasie et al., 2006) and potato (Souza-Dias et al., 2008) and is only rarely found in non-cultivated plants such as Nicandra physaloides (family Solanaceae) (Barbosa et al., 2009). Over a 3 year period we obtained more than 50 full-length DNA-A sequences of each of these viruses isolated from a mixture of crops and non-cultivated plants. We compared the molecular evolution and population genetics of the mostly weed-infecting MaYSV to the predominantly tomato-infecting ToSRV. Our results bolster the assertion that infection of indigenous and non-cultivated plant species leads to higher levels of standing genetic variability, apparently driven by higher levels of detectable recombination.

**RESULTS**

**Natural infection of the crop plant Phaseolus vulgaris by MaYSV and of the wild host Sida spp. by ToSRV**

MaYSV was originally described infecting leguminous non-cultivated hosts (Macroptilium lathyroides, Calopogonium mucunoides and Canavalia sp.) (Silva et al., 2012). However, in 2011 MaYSV was readily isolated from both M. lathyroides and a crop plant, P. vulgaris (common bean), in Alagoas. Forty-four DNA-A components were cloned from these samples (Table S1, available in JGV Online). Pairwise comparisons of the DNA-A sequences (data not shown) revealed that only one begomovirus was present, with 90–99 % identity to MaYSV. Therefore, based on the criteria established by the Geminiviridae Study Group of the ICTV (Brown et al., 2012) these genomic components represent additional isolates of this species. MaYSV appeared to be the prevalent begomovirus infecting common bean in this portion of Alagoas, since Bean golden mosaic virus was not detected in P. vulgaris in our study. Overall, 56 whole genomes were included in the MaYSV dataset, with 23 (41 %) isolated from non-cultivated hosts.

Despite sampling many non-cultivated species neighbouring infected tomato plants, we were only able to find ToSRV in two samples of the malvaceous non-cultivated host, Sida sp. Therefore, 53 out of 55 whole ToSRV genomes were isolated from tomatoes and only two from non-cultivated hosts.

**MaYSV population is more variable than the ToSRV population**

Although the ToSRV and MaYSV populations were sampled in similarly sized geographical areas, they differed profoundly in their levels of standing genetic variability. While the average pairwise number of nucleotide differences (π) for the full-length DNA-A of the ToSRV population was 0.0084 (Table 1), this same statistic was about eightfold higher for the MaYSV population (π=0.0658). Interestingly, the variability within the MaYSV population was not evenly distributed throughout the genome. The MaYSV Rep gene was 3.8 times more variable than its CP gene, and more than 10-fold higher than the ToSRV Rep (Table 1). Within the MaYSV Rep, the N-terminal half was much more variable than the C-terminal half – again, a distribution of variation...
Table 1. Genetic variability of the begomoviruses MaYSV and ToSRV

<table>
<thead>
<tr>
<th>Population</th>
<th>No. sequences</th>
<th>DNA-A H_d</th>
<th>DNA-A π</th>
<th>CP π</th>
<th>Rep π</th>
</tr>
</thead>
<tbody>
<tr>
<td>ToSRV (total)</td>
<td>55</td>
<td>0.999 (±0.004)</td>
<td>0.00844 (±0.00136)</td>
<td>0.00963 (±0.00238)</td>
<td>0.00743 (±0.00140)</td>
</tr>
<tr>
<td>MaYSV (total)</td>
<td>56</td>
<td>1.000 (±0.003)</td>
<td>0.06580 (±0.00231)</td>
<td>0.02852 (±0.00322)</td>
<td>0.10889 (±0.00456)</td>
</tr>
<tr>
<td>P. vulgaris</td>
<td>33</td>
<td>1.000 (±0.007)</td>
<td>0.06264 (±0.00300)</td>
<td>0.02189 (±0.00102)</td>
<td>0.11175 (±0.00599)</td>
</tr>
<tr>
<td>Non-cultivated hosts</td>
<td>23</td>
<td>1.000 (±0.013)</td>
<td>0.06702 (±0.00425)</td>
<td>0.03616 (±0.00670)</td>
<td>0.10547 (±0.00882)</td>
</tr>
</tbody>
</table>

not shared by the ToSRV Rep (Fig. 1). The nucleotide diversity of the ToSRV CP gene was similar to that of MaYSV.

We also compared the variability between MaYSV isolates sampled from crops and non-cultivated hosts. Each MaYSV subset was still markedly more variable than the whole ToSRV population, with isolates sampled from P. vulgaris and non-cultivated hosts showing similar levels of standing genetic variability (π = 0.0626, SD = 0.003 and π = 0.0670, SD = 0.0043, respectively).

Phylogenetic analysis

The ToSRV CP and Rep ML phylogenetic trees are not congruent (Fig. 2), but the topological differences between them were due to a strongly supported recombination event in the CP gene of ToSRV-[BR:Vic20:10] (P = 2.18 × 10⁻¹⁰, Table 2). Without this isolate, the clustering of isolates in the CP and Rep trees are identical (Fig. S1), mimicking the clustering found in the Rep tree on Fig. 2. Both ML trees showed strong support for isolates from Florestal being a separate population from those from Carandai, Jaiba and Vicosa (Fig. 2). There is rough clustering of isolates from each sampling site together, with Jaiba being nested within isolates from Vicosa. However, the lack of bootstrap support for geographical structure among these three sites (and the weakly supported clustering of Vic06 and Vic07 with isolates from Carandai) suggests that migration may occur between these subpopulations. Consistent with our phylogenetic analyses, Wright’s fixation index F, based on the ToSRV CP dataset, indicated genetic differentiation amongst isolates sampled from different geographical locations (FST = 0.51066).

The MaYSV CP and Rep trees were highly incongruent (Fig. 3). The MaYSV Rep tree shows four clades with significant genetic distance between them (two with 100 % bootstrap support), and the CP shows fewer well-supported clades and less genetic variability among isolates. Both trees showed little evidence of geographical structuring. In fact, Wright’s fixation index F based on the MaYSV CP dataset indicated less evidence of genetic differentiation (FST = 0.11428) than the ToSRV population.

MaYSV isolates have a mixture of different recombinant patterns

While ToSRV showed little evidence that recombination has significantly contributed to its evolution (Table 2), MaYSV showed both inter- and intraspecific recombination events. Corroborating the conflicting MaYSV CP and Rep phylogenies, we identified a total of six unique potential recombination events in the MaYSV population. Most events involved breakpoints located inside the Rep gene and in the common region (events 1, 2, 3, 4 and 6; Table 2), though one event had breakpoints within the CP gene (event 5; Table 2). Isolates that showed evidence of a
shared recombination event in the Rep sequence were readily identified in well supported clades observed on the Rep ML phylogenetic tree (Fig. 3). There were no strong geographical patterns associated with recombination events. For instance, four out of five recombination events detected in MaYSV Rep (events 1, 2, 4 and 6) were detected in isolates sampled in Olho d’A´ gua das Flores, but two of these events (events 1 and 6) were observed elsewhere.

Adaptive selection does not explain the higher begomovirus variability in non-cultivated hosts

We investigated the extent that positive and negative selection at the amino acid level had shaped the standing genetic variability in the CP and Rep datasets (comprised of all isolates obtained from cultivated and non-cultivated hosts) of both viruses. All datasets showed $d_S/d_K$ ratios ($\omega$) lower than 1 (Table 3), indicating negative selection. However, the wide variation of the values indicated that each gene/population might be under different selective constraints. The $\omega$ value for the ToSRV CP dataset (0.446493) was considerably higher than that observed for the MaYSV CP (0.0514088). The high $\omega$ of the ToSRV CP was observed even with the elimination of the single recombinant isolate, Vic20 ($\omega=0.27438$). The $\omega$ value for the ToSRV Rep (0.268859) was only slightly higher than that observed for the MaYSV Rep dataset (0.208195).

![Fig. 2. Midpoint-rooted ML trees based on the CP (a) and Rep (b) nucleotide sequences of isolates from the ToSRV population. Nodes with bootstrap values equal or higher than 80 % are indicated by filled circles, and those with values lower than 80 and higher than 50 % by empty circles. The contribution of parental sequences to the unique recombination event detected within the CP coding sequence (event 1, isolate Vic20; Table 2) is shown as a diagram above the branch where the substitutions due to recombination were mapped (in red colour). Isolates sampled from non-cultivated hosts are shown in green.](http://vir.sgmjournals.org)
Table 2. Recombination events detected by RDP in the DNA-A of ToSRV and MaYSV populations

<table>
<thead>
<tr>
<th>Event</th>
<th>Recombinant</th>
<th>Recombination breakpoints</th>
<th>Parents</th>
<th>Methods†</th>
<th>P-value‡</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Initial</td>
<td>Final</td>
<td>Major</td>
<td>Minor</td>
</tr>
<tr>
<td>ToSRV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Vic20:10</td>
<td>537</td>
<td>1011</td>
<td>Unknown</td>
<td>HV25:10</td>
</tr>
<tr>
<td>MaYSV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Oaf12:11</td>
<td>2654</td>
<td>1661</td>
<td>Unknown</td>
<td>Pdi1:10</td>
</tr>
<tr>
<td></td>
<td>Bat1:10</td>
<td></td>
<td></td>
<td>Unknown</td>
<td>Oaf1:10</td>
</tr>
<tr>
<td></td>
<td>Inp1:10</td>
<td></td>
<td></td>
<td>Unknown</td>
<td>Oaf2:10</td>
</tr>
<tr>
<td></td>
<td>Pir1:10</td>
<td></td>
<td></td>
<td>Unknown</td>
<td>Ced1:09</td>
</tr>
<tr>
<td></td>
<td>Crb8:11</td>
<td></td>
<td></td>
<td>Deg1:10</td>
<td>Ced1:09</td>
</tr>
<tr>
<td></td>
<td>Crb10:11</td>
<td></td>
<td></td>
<td>Inp2:10</td>
<td>Ced1:09</td>
</tr>
<tr>
<td></td>
<td>Crb13:11</td>
<td></td>
<td></td>
<td>Mac5:10</td>
<td>Ced1:09</td>
</tr>
<tr>
<td></td>
<td>Crb14:11</td>
<td></td>
<td></td>
<td>Crb11:11</td>
<td>Ced1:09</td>
</tr>
<tr>
<td></td>
<td>Crb15:11</td>
<td></td>
<td></td>
<td>Crb12:11</td>
<td>Ced1:09</td>
</tr>
<tr>
<td></td>
<td>Crb16:11</td>
<td></td>
<td></td>
<td>Crb17:11</td>
<td>Ced1:09</td>
</tr>
<tr>
<td></td>
<td>Oaf16:11</td>
<td></td>
<td></td>
<td>Crb18:11</td>
<td>Ced1:09</td>
</tr>
<tr>
<td></td>
<td>Oaf17:11</td>
<td></td>
<td></td>
<td>Oaf15:11</td>
<td>Ced1:09</td>
</tr>
<tr>
<td></td>
<td>Oaf20:11</td>
<td></td>
<td></td>
<td>Oaf18:11</td>
<td>Ced1:09</td>
</tr>
<tr>
<td></td>
<td>Oaf21:11</td>
<td></td>
<td></td>
<td>Oaf19:11</td>
<td>Ced1:09</td>
</tr>
<tr>
<td></td>
<td>Oaf22:11</td>
<td></td>
<td></td>
<td>Crb1:11</td>
<td>Ced1:09</td>
</tr>
<tr>
<td></td>
<td>Oaf23:11</td>
<td></td>
<td></td>
<td>Crb2:11</td>
<td>Ced1:09</td>
</tr>
<tr>
<td></td>
<td>Oaf3:11</td>
<td></td>
<td></td>
<td>Oaf6:11</td>
<td>Ced1:09</td>
</tr>
<tr>
<td></td>
<td>Oaf19:11</td>
<td></td>
<td></td>
<td>Oaf1:10</td>
<td>Bat1:10</td>
</tr>
<tr>
<td></td>
<td>Oaf6:11</td>
<td></td>
<td></td>
<td>Oaf2:10</td>
<td>Inp1:10</td>
</tr>
<tr>
<td></td>
<td>Oaf9:11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Inp1:10</td>
<td>1984</td>
<td>2648</td>
<td>Oaf14:11</td>
<td>Crb8:11</td>
</tr>
<tr>
<td>4</td>
<td>Oaf25:11</td>
<td>1798</td>
<td>(?) 2653</td>
<td>Oaf24:11</td>
<td>Unknown</td>
</tr>
<tr>
<td></td>
<td>Oaf22:11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Oaf7:11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Oaf8:11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Oaf10:11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Oaf11:11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Oaf12:11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Although GARD was unable to detect the well-supported recombination event in the ToSRV CP dataset, no sites were under statistically significant negative selection, including or excluding Vic20 (Fig. 4 and Fig. S2). Excluding the recombinant Vic20 from the dataset, Random Effects Likelihood (REL) detected 20 sites under positive selection. However, the evidence for selection on individual sites is weak, as no positively selected sites were detected using the Single Likelihood Ancestor Counting (SLAC) or Partitioning for Robust Inference of Selection (PARRIS) methods. Two negatively selected sites were detected by the SLAC method in the ToSRV Rep dataset (Fig. 4). In addition, nine positively selected sites were detected by REL (Fig. S2), with six of them located between motif III in the catalytic domain and motif Walker A in the ATPase domain. PARRIS did not identify any sites under positive selection.

Recombinant partitions were readily detected by GARD in the MaYSV CP and Rep datasets. Nevertheless, a higher number of sites were under detectable selection in both MaYSV genes. In the CP dataset, all codons were identified as being under negative selection by REL (Fig. S2). In the MaYSV Rep, 106 sites were shown to be under negative selection by SLAC or REL, and 12 sites were under positive selection by REL (Fig. 4, Fig. S2). Many sites in the highly variable Rep N-terminal showed high synonymous substitution rates with statistical evidence of negative selection (Fig. 4). Although GARD was able to detect recombination events in this dataset, none of the positively selected sites detected by REL correlated with the host from which the isolates were obtained (non-cultivated vs cultivated), but were clearly related to a given recombinant event, indicating that they most probably represent a spurious selection signal due to recombination. Results obtained by PARRIS confirmed the absence of sites under positive selection.

Together, our results indicate that adaptive selection at the amino acid level is not driving the higher variability in the MaYSV population.

Relative contribution of mutation and recombination to the genetic variability of ToSRV and MaYSV populations

We mapped all substitutions over the branches in our midpoint-rooted ML trees and determined which branches were associated with well-supported recombination events. Out of a total of 99 substitutions, 42 substitutions were associated with well-supported recombination events in the ToSRV CP phylogeny, the single recombination event detected in the ToSRV Rep dataset, and all of the substitutions along the phylogeny are presumably due to mutation. As there were no detectable recombination events in the ToSRV Rep dataset, all the substitutions along the phylogeny are presumably due to mutation.

We mapped all substitutions over the branches in our midpoint-rooted ML trees and determined which branches were associated with well-supported recombination events. Out of a total of 99 substitutions, 42 substitutions were associated with well-supported recombination events. As there were no detectable recombination events in the ToSRV Rep dataset, all of the substitutions along the phylogeny are presumably due to mutation. Out of a total of 99 substitutions, 42 substitutions were associated with well-supported recombination events in the ToSRV CP phylogeny, the single recombination event detected in the ToSRV Rep dataset, and all of the substitutions along the phylogeny are presumably due to mutation. As there were no detectable recombination events in the ToSRV Rep dataset, all of the substitutions along the phylogeny are presumably due to mutation.

Table 2. cont.

<table>
<thead>
<tr>
<th>Event</th>
<th>Recombinant</th>
<th>Recombination breakpoints</th>
<th>Parents</th>
<th>Methods†</th>
<th>P-value‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Oaf1:11</td>
<td>415</td>
<td>Crb1:11</td>
<td>GBM3</td>
<td>1.09 × 10⁻⁶</td>
</tr>
<tr>
<td>6</td>
<td>Prl:10</td>
<td>(?) 2159</td>
<td>Unknown</td>
<td>MCS</td>
<td>1.39 × 10⁻²⁰</td>
</tr>
<tr>
<td></td>
<td>Bat:11</td>
<td>(?) 2653</td>
<td>Oaf1:11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Inp:10</td>
<td></td>
<td>Oaf2:10</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Crb8:11</td>
<td></td>
<td>Oaf4:11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Crb9:11</td>
<td></td>
<td>Oaf5:11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Crb10:11</td>
<td></td>
<td>Oaf6:11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Crb11:11</td>
<td></td>
<td>Oaf7:11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Crb12:11</td>
<td></td>
<td>Oaf8:11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Crb13:11</td>
<td></td>
<td>Oaf9:11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Crb14:11</td>
<td></td>
<td>Oaf10:11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Oaf16:11</td>
<td></td>
<td>Oaf11:11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Oaf17:11</td>
<td></td>
<td>Oaf12:11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Oaf18:11</td>
<td></td>
<td>Oaf13:11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Oaf19:11</td>
<td></td>
<td>Oaf14:11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Oaf20:11</td>
<td></td>
<td>Oaf15:11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Oaf21:11</td>
<td></td>
<td>Oaf16:11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Oaf22:11</td>
<td></td>
<td>Oaf17:11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Oaf23:11</td>
<td></td>
<td>Oaf18:11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Oaf24:11</td>
<td></td>
<td>Oaf19:11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Oaf25:11</td>
<td></td>
<td>Oaf20:11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Oaf26:11</td>
<td></td>
<td>Oaf21:11</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Numbering starts at the first nucleotide after the cleavage site at the origin of replication and increases clockwise. (?) Indicates that the breakpoint could not be precisely pinpointed.
†R, RDP; G, GeneConv; B, Bootscan; M, MaxChi; C, Chimaera; S, SisScan; 3, 3Seq.
‡The reported P-value is for the program in bold, underlined type and is the lowest P-value calculated for the event in question.

http://vir.sgmjournals.org
region (spanning about 75% of the CP gene) on a single branch of the tree. In this considerably more variable dataset, recombination accounted for only 16% of the substitutions (Table 4).

As expected, a considerably higher number of substitutions was observed for the MaYSV Rep ML tree: 553 substitutions. The unique recombination events 2, 4 and 6 were readily assigned to the long branches (the brown, red and yellow branches, respectively, in Fig. 3b), with another long branch leading to the clade containing events 4 and 6 (the orange branch; Fig. 3b). For events 2 (shared by isolates

![Fig. 3.](image-url)
Oaf6, Oaf9 and Oaf19) and 4 (shared by all isolates in clade II), 22 and 36 substitutions were counted as associated to them, respectively. Event 6 (shared by all 23 isolates in clade I) accounted for the largest individual contribution amongst the unique recombination events: 37 substitutions. Additionally, 16 changes on the branch leading to the clades associated with events 4 and 6 were either in the region of overlap between events 4 and 6, or within the larger event 6, and were also assigned to \( \eta_{\text{recombination}} \) (\( \eta_r \)) (Table 4). Event 3 (assigned to the purple terminal branch) was exclusively detected in the Inp1 isolate and a total of 12 substitutions were counted in association with this event.

Although event 1 was well supported by the methods of analysis contained in RDP, the exact location of the breakpoint was variable and frequently ended in the intergenic region instead of the Rep gene. Only those substitutions in the terminal branches leading to sequences whose end breakpoint of this event occurred inside the Rep gene (isolates Oaf3, Oaf7, Oaf8, Oaf11, Oaf12, Oaf16, Oaf21, Oaf24, Oaf25, Crb10 and Inp1) were added to the \( \eta_r \). A total of five substitutions could be associated with this event.

Overall, 201 substitutions were counted as due to recombination over the MaYSV Rep phylogeny, which represented a relative contribution of 36.3% to the standing genetic variability (Table 4).

**DISCUSSION**

Most studies on the diversity of begomovirus populations have focused on species diversity (Ala-Poikela *et al.*, 2005; Bull *et al.*, 2006; Fernandes *et al.*, 2008; García-Andrés *et al.*, 2006; Lefèvre *et al.*, 2007b; Lozano *et al.*, 2009; Ndunguru *et al.*, 2005; Reddy *et al.*, 2005; Ribeiro *et al.*, 2003; Rothenstein *et al.*, 2006; Sserubombwe *et al.*, 2008). In

**Table 4.** Relative contribution of mutation and recombination to the standing genetic variability of ToSRV and MaYSV populations

<table>
<thead>
<tr>
<th>Population</th>
<th>CP</th>
<th>Rep</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \eta_r )</td>
<td>( \eta_{\mu} )</td>
</tr>
<tr>
<td>ToSRV</td>
<td>42 (42.4%)</td>
<td>57 (57.6%)</td>
</tr>
<tr>
<td>MaYSV</td>
<td>49 (15.7%)</td>
<td>264 (84.3%)</td>
</tr>
</tbody>
</table>

**Fig. 4.** Difference between the non-synonymous and synonymous substitution rates \( (d_N - d_S) \) calculated across the codons of the CP (a, c) and Rep sequences (b, d) of isolates from the ToSRV (a, b) and MaYSV (c, d) populations using the SLAC method. Sites with statistical evidence of negative selection are indicated by blue bars, while neutrally evolving sites are indicated by black bars.
contrast, little is known about the standing genetic variability within species, especially in begomovirus populations in non-cultivated hosts. Here, we present a study contrasting the molecular variability of begomovirus populations of the mostly weed-infecting MaYSV and the predominantly tomato-infecting ToSRV, and discuss the relative contribution of evolutionary processes on this variability. Our results, based on 111 viral genome sequences cloned from samples collected over a 3 year period, support the hypothesis that although the genetic structure of viral populations is modulated by the common processes of mutation, recombination, selection and genetic drift (García-Arenal et al., 2003; Roossinck, 2003), the relative contribution of each process can vary widely among begomovirus populations.

The MaYSV population was notably more variable than the ToSRV population, largely due to a number of recombination events in the Rep gene. Even MaYSV isolates sampled from different events in the Rep gene. Even MaYSV isolates sampled from the crop *P. vulgaris* (collected from two locations, Craibas and Olho d’Água das Flores) were considerably more variable than all ToSRV isolates sampled from tomato plants (collected from four different locations). Previous studies have indicated that begomovirus populations infecting non-cultivated hosts are more variable than crop-infecting ones (Rocha, 2011; Silva et al., 2012). The standing genetic variabilities of four tomato-infecting begomovirus populations were much lower than that of a begomovirus population from a non-cultivated host (*Blainvillea yellow spot virus*) sampled in south-eastern Brazil (although these results are based on different sample sizes) (Rocha, 2011). The genetic structure was also determined for a population of *Cleome leaf crumple virus* infecting an annual weed (*Cleome affinis*, family Capparaceae) often associated with leguminous crops in Brazil (Silva et al., 2011). Several recombination events and a high molecular variability were estimated for this population.

Although an increasing body of evidence points to a higher genetic variability in begomovirus populations infecting non-cultivated hosts, all previous studies have detected whether various evolutionary processes have contributed to the population genetic variability without attempting to determine their relative contribution to the total variability in these populations. Our novel substitution counting method allowed us to assess the relative contribution of recombination and mutation to the standing genetic variability of two begomoviruses – a step towards disambiguating the effects of various evolutionary processes on viral genetic variation.

**Recombination and the genetic structure of begomovirus populations**

Recombination is the most studied evolutionary process acting on geminivirus populations and has greatly contributed to their evolution (Briddon et al., 1996; García-Andrés et al., 2007a; Garcia-Andrés et al., 2007b; Padidam et al., 1999; Pita et al., 2001; Schnippenkoetter et al., 2001). In begomoviruses, the non-random location of recombination breakpoints is conserved amongst mono- and bipartite genomes, with hot spots in the Rep N-terminal portion and in the 5’-end of the common region (Lefeuvre et al., 2007a, b; Prasanna & Rai, 2007). Our analyses of the MaYSV population indicated that most unique recombination events involved the Rep gene, with at least one breakpoint located in the N-terminal portion and/or the 5’-end of the common region. In addition, we also observed that the N-terminal portion accounted for the largest fraction of the total variability in the Rep gene, suggesting that recombination may be the evolutionary process responsible for this uneven distribution of polymorphisms.

It has been shown that recombination events that preserve co-evolved intragenome interactions (protein–protein and/or protein–DNA) are favoured by selection (Martin et al., 2011). In fact, the linkage between the Rep N-terminal portion and the 5’ common region was not broken by recombination events detected in the MaYSV population. These portions were exchanged as blocks that preserve them as originating from the same genetic background. This could increase the frequency in which viable recombinants for the Rep sequence are maintained by preserving the interaction between the catalytic domains (in the Rep N-terminal portion) and their cognate iterative elements located in the 5’ portion of the common region (Martin et al., 2011). Although a reduced number of breakpoints occur within the CP-encoding sequence, most of them are located in the central portion of the gene (Lefeuvre et al., 2007b). In agreement, two unique events (the single event detected in the ToSRV population and the MaYSV event 5) had initial breakpoints located in the central portion of the gene and spanned its entire C-terminal portion.

**Purifying selection in viral genes**

Although previous studies indicate that the main type of selection acting on the CP and Rep genes in begomovirus populations is purifying selection (García-Andrés et al., 2007a; Sanz et al., 1999; Silva et al., 2011, 2012), we also assessed the possible contribution of adaptive selection in shaping the standing genetic variability in the ToSRV and MaYSV populations, since a small fraction of codons could be under different selective constraints. Interestingly, the \( d_\beta/d_\alpha \) ratio estimated for the ToSRV CP was markedly high (\( \omega = 0.446 \)), even after the exclusion of the recombinant isolate Vic20 (\( \omega = 0.274 \)). These values are high even when compared with non-vector-borne ssRNA viruses such as *Prune dwarf virus* (family Bromoviridae; \( \omega = 0.222 \)) and similar to *Pepper mild mottle virus* (genus Tobamovirus; \( \omega = 0.301 \)) (Chare & Holmes, 2004). Despite the lower molecular variability, REL detected stronger evidence of adaptive selection in the CP (two sites) and especially in the ToSRV Rep (nine sites). Interestingly, out of nine positively selected sites in the ToSRV Rep, six were located between motif III in the catalytic domain and motif Walker
A in the ATPase domain. This region is putatively involved in Rep oligomerization and interaction with host factors (Hanley-Bowdoin et al., 2004; Orozco et al., 2000). Although REL is considerably less conservative than SLAC in detecting positively selected sites, it is important to note that the Rep ToSRV dataset was free of any detectable evidence of recombination that is known to affect the results of selection analyses.

In striking contrast, we did not detect any evidence of adaptive selection acting on both MaYSV genes, indicating that adaptive selection did not contribute to the higher molecular variability in the MaYSV population.

Although a higher variability was found in the Rep N-terminal portion (probably due to recombination) it was composed mostly of synonymous substitutions. The Rep catalytic domain is located in the N-terminal portion and includes three motifs which are conserved in proteins involved in RCR (motif I: FLTY; motif II: HxH; and motif III: YxxxV) (Ilyina & Koonin, 1992). Adjacent to these elements are the DNA-binding specificity determinants involved in high affinity DNA-binding (Londoño et al., 2010). The integrity of these elements in terms of amino acid sequence and structure is important for viral replication and therefore suggests that although the nucleotide sequence in the 5’ portion of the Rep gene may harbour a high variability, most of these substitutions would tend to preserve the amino acid sequence and consequently protein structure as well.

**Host preference and recombination**

The lower synonymous site variation in ToSRV population suggests that this population may have been recently introduced in the sampled geographical areas (either from imported tomatoes or from a wild host) and/or, less likely, experienced a recent genetic bottleneck. Low levels of molecular variability were also observed in genetically differentiated subpopulations of begomoviruses causing TYLCD in Spain and Italy. This scenario was consistent with the foundation of these subpopulations by few variants of limited genetic variability (García-Andrés et al., 2007a). It is possible that the lack of alternative perennial hosts that would maintain high ToSRV effective population sizes between cultivation periods results in a strong reduction of variability. MaYSV is able to efficiently infect both *P. vulgaris* and perennial leguminous non-cultivated hosts. These latter hosts could represent an important epidemiological component for the MaYSV population for allowing that genetic variability is maintained between cultivation periods. Furthermore, by infecting new hosts these populations may be subject to environments inaccessible to viruses that have a narrow host range (Power, 2000). As a consequence, these viruses could experience high frequencies of interspecific recombination, resulting in a rapid evolution of these populations. Studies attempting to assess the genetic variability of begomovirus populations on *Solanum nigrum* (an indigenous weed in Spain) showed that this host is an efficient reservoir of all previously characterized begomoviruses species and strains causing TYLCD (García-Andrés et al., 2006). Additionally, an uncharacterized recombinant isolate was detected in this host. In fact, several novel species of begomoviruses causing TYLCD are actually interspecific recombinants (Monci et al., 2002; Navas-Castillo et al., 2000). Begomoviruses causing CMD are the major biotic constraint to cassava cultivation in Africa (Legg & Fauquet, 2004; Legg & Thresh, 2000; Ndunguru et al., 2005; Were et al., 2004), and multiple natural reservoirs for begomoviruses causing CMD have been described previously (Alabi et al., 2007, 2008). Associated with the ability to infect many alternative hosts, these viruses appear to evolve largely by intra- and interspecific recombination (Berrie et al., 2001; Fondong et al., 2000; Pita et al., 2001; Tiendrébéogo et al., 2012; Zhou et al., 1997).

**Mutation accounts for most begomovirus genetic variability**

Our results suggest that recombination explains the higher molecular variability of the MaYSV population compared with the ToSRV population. However, both populations were dominated by mutational diversification. By mapping the recombination events onto ML trees, we were able to distinguish the individual contributions associated with these two mechanisms that create variability. Then, we took advantage of the additive character of η to express the individual contributions of these processes as fractions of the total variability ($\eta^{total} = \eta_r + \eta_m$). This simple statistic is not without its flaws, chiefly that recombination is known to reduce the accuracy of the phylogenetic trees upon which this method relies (Posada & Crandall, 2002). However, phylogeny independent measures of population variability, such as pairwise comparisons, are non-additive, which complicates partitioning variability due to recombination and mutation.

Some of the variation that our method attributes to mutation is undoubtedly due to recombination that was not statistically detectable. Therefore, our estimates should be interpreted as the minimal relative contribution of recombination. Furthermore, while our method attributes 100 % of changes on specified branches within identified recombination breakpoints to recombination, this is not a liberal assumption. The reconstructed sequence of the recombinant region that finds each recombinant clade may not be the correct ancestral state, but it is unlikely to contain dramatically more substitutions than the actual first recombinant. All subsequent changes that occur in that region on branches to descendant nodes are counted as mutations.

Using this novel partitioning, we assessed recombination as the source of up to 50 % of the variation in begomovirus populations. Importantly, our results indicate that its relative contribution to the total variability does not necessarily correlate with the number of detectable recombination events: the highest relative contribution of
recombination was in the ToSRV CP, which was the least variable dataset analysed. On the other hand, results from the MaYSV Rep indicate that in recombination hotspots, recombination may be at least as important as mutation in generating variability.

We have assessed the molecular variability of two Brazilian begomoviruses populations, and found that in certain portions of the genome recombination could be responsible for a considerable fraction of the total variability in addition to mutation. Additionally, we also found evidence that the distinct evolutionary patterns might, at least in part, be related to the ability to infect non-cultivated, perennial hosts. Such ability would thus be an important factor in the rapid evolution of these populations.

**METHODS**

**Sequence datasets.** A total of 55 full-length DNA-A sequences corresponding to ToSRV and 56 corresponding to MaYSV were analysed [GenBank accession numbers: JN419005, JN419007, JN419009, JN419012–JN419016, JN419018–JN419020, JN419022 and KC004091–KC004134 (MaYSV), KC004068–KC004090 and JX865615–JX865650 (ToSRV)]. The ToSRV dataset comprised isolates obtained from tomato (53 isolates) and Sida spp. plants (two isolates) collected in three different municipalities in the state of Minas Gerais (MG) from 2008 to 2010 (Table S1; González-Aguilera et al., 2012). Twelve of the MaYSV sequences were isolated from leguminous non-cultivated hosts collected at nine municipalities of the states of Alagoas (AL), Sergipe (SE) and Paraíba (PB) in 2009 and 2010 as described previously (Silva et al., 2012), and 44 additional sequences were obtained from M. lathyroides and the crop P. vulgaris (common bean) samples showing typical symptoms of begomovirus infection collected in Olho d’Água das Flores and Craibas (AL) in 2011 (Table S1). The sampling areas in MG and AL/SE/PB have similar sizes (approximately 45 000 km²).

Total DNA was extracted from fresh tissue or herbarium-like (pressed and dried) samples as described by Doyle & Doyle (1987). Full-length viral genomes were amplified by rolling-circle amplification, according to the standard circular DNA cloning method (Inoue-Nagata et al., 2004). Single genome-length fragments were excised from these concatamers with BamHI and ligated into the pBlUESCRIPT-KS+ (Stratagene) plasmid vector, previously cleaved with the same enzyme. Viral inserts were commercially sequenced (Macrogen Inc.) by primer walking. All genome sequences were organized to begin at the nicking site in the invariant nonanucleotide at the origin of replication (5’-TAATATT/AC-3’).

**Multiple sequence alignments and phylogenetic analysis.** Multiple sequence alignments were prepared for the full-length DNA-A and for the CP and Rep coding sequences of each viral species (ToSRV and MaYSV) using MUSCLE (Edgar, 2004). ML trees were inferred for CP and Rep sequences using PAUP* v. 4.0 (Swofford, 2003). These genes were chosen due to their essential role for insect transmission and viral replication, respectively, but they encompass about 70% of the full-length DNA-A genome. The best-fit model of nucleotide substitution was determined using MODELTEST (Posada & Crandall, 1998) by the Akaike Information Criterion (Table S2). The heuristic ML search was initiated with a neighbour-joining tree, with optimization using the tree-bisection-reconnection algorithm. The robustness of each individual branch was estimated from 2000 nearest neighbour interchange bootstrap replicates. Trees were visualized and edited using FigTree (tree.bio.ed.ac.uk/software/figtree).

**Variability indices.** The haplotype diversity (Hd), the mean pairwise number of nucleotide differences per site (nucleotide diversity, π) and Wright’s F fixation index were calculated using DnaSP v. 5.10 (Rozas et al., 2003). The π statistic was also calculated on a sliding window of 100 bases, with a step size of 10 bases in order to estimate the nucleotide diversity across the length of the CP and Rep datasets.

**Recombination analysis.** Recombination analysis was performed using the RDP, Geneconv, Bootscan, Maximum Chi Square, Chimaera, SisterScan and 3Seq methods as implemented in Recombination Detection Program (RDP) version 3.0 (Martin et al., 2010) (RDP project files available from the authors upon request). Alignments were scanned with default settings for the different methods. Statistical significance was inferred by P-values lower than a Bonferroni-corrected cut-off. Only recombination events detected by at least three of the analysis methods available in the program were considered reliable.

**Detection of positive and negative selection at amino acid sites.** To detect sites under positive and negative selection, we analysed the CP and Rep datasets using three ML-based methods implemented in DataMonkey (www.datamonkey.org): SLAC, REL and PARRIS (Kosakovsky Pond & Frost, 2005; Scheffler et al., 2006). As recombinant sequences cause spurious selection results, we searched for breakpoints in sequences implicated as recombinant by GARD prior to running these analyses. Additionally, PARRIS allowed synonymous rate variation, topology and branch lengths to vary across recombination breakpoints. The SLAC method was also used to estimate the mean ratios of non-synonymous to synonymous substitutions (dN/dS) for the CP and Rep datasets from both begomovirus populations based on the inferred GARD phylogenetic trees. These methods were applied under the nucleotide substitution models determined in MODELTEST as described in Table S2. Bayes factors larger than 50 and P-values smaller than 0.01 were used as thresholds for the REL method.

**Relative contribution of recombination and mutation in the genetic variability of ToSRV and MaYSV populations.** A phylogeny based approach was developed to determine the relative contribution of recombination and mutation to the total variability observed in the ToSRV and MaYSV populations. We identified groups of sequences descended from a shared recombination event (based on RDP analysis), and frequently these formed clades on midpoint-rooted CP and Rep ML trees. The sequence of the ancestral node of each of these clades reflected the recombination event (ancestral state reconstruction by marginal ML method in PAUP*), but its parental node did not. Wherever possible, we assigned each recombination event to the branch between these nodes. We did not consider the minor/major parent information from RDP analyses to be definitive, and instead considered the portion of the sequence that differed the most from the parental nodes on the ML tree to be the recombined, introduced block. In the case of recombination events associated with only one sequence, the event was assigned to the branch leading to the corresponding tip. Using PAUP*, we identified all substitutions that occurred over each ML tree, and calculated η, the total number of substitutions over the phylogeny (Fu & Li, 1993). We then looked at the location within the gene of substitutions that occurred on branches associated with recombination. Substitutions that occurred in the region likely introduced by recombination were added to η, the total number of substitutions on the phylogeny likely due to recombination. All other substitutions, including those on branches associated with recombination but outside of the region implicated in the recombinational event, were added to ηmutation (ηm). Thus, η = ηi + ηm.

**ACKNOWLEDGEMENTS**

The authors wish to thank Eduardo Mizubuti and Elizabeth Fontes for their continuous and invaluable support. A.T.M.L. was the...
recipient of a CNPq doctoral fellowship. S.D. was supported by the NSF (DEB 1026095). F. M. Z. is supported by CNPq (484447/2011-4) and Fapemig (APQ-949-09).

REFERENCES


A T. M. Lima and others


