Begomovirus ‘melting pot’ in the south-west Indian Ocean islands: molecular diversity and evolution through recombination

P. Lefeuvre,1 D. P. Martin,2 M. Hoareau,1 F. Naze,1 H. Delatte,1 M. Thierry,1 A. Varsani,3 N. Becker,4 B. Reynaud1 and J.-M. Lett1

Correspondence
J.-M. Lett
lett@cirad.fr

1CIRAD, UMR 53 PVBM CIRAD-Université de la Réunion, Pôle de Protection des Plantes, 7 Chemin de l’IRAT, 97410 Saint Pierre, La Réunion, France
2Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, Observatory 7925, South Africa
3Electron Microscopy Unit, University of Cape Town, Rondebosch 7701, South Africa
4Museum National d’Histoire Naturelle, Dept RDDM, USM 501, CNRS UMR 5166, Evolution des Régulations Endocriniennes, 57 rue Cuvier, CP 32, 75005 Paris, France

Received 20 June 2007
Accepted 9 August 2007

During the last few decades, many virus species have emerged, often forming dynamic complexes within which viruses share common hosts and rampantly exchange genetic material through recombination. Begomovirus species complexes are common and represent serious agricultural threats. Characterization of species complex diversity has substantially contributed to our understanding of both begomovirus evolution, and the ecological and epidemiological processes involved in the emergence of new viral pathogens. To date, the only extensively studied emergent African begomovirus species complex is that responsible for cassava mosaic disease. Here we present a study of another emerging begomovirus species complex which is associated with serious disease outbreaks in bean, tobacco and tomato on the south-west Indian Ocean (SWIO) islands off the coast of Africa. On the basis of 14 new complete DNA-A sequences, we describe seven new island monopartite begomovirus species, suggesting the presence of an extraordinary diversity of begomovirus in the SWIO islands. Phylogenetic analyses of these sequences reveal a close relationship between monopartite and bipartite African begomoviruses, supporting the hypothesis that either bipartite African begomoviruses have captured B components from other bipartite viruses, or there have been multiple B-component losses amongst SWIO virus progenitors. Moreover, we present evidence that detectable recombination events amongst African, Mediterranean and SWIO begomoviruses, while substantially contributing to their diversity, have not occurred randomly throughout their genomes. We provide the first statistical support for three recombination hot-spots (V1/C3 interface, C1 centre and the entire IR) and two recombination cold-spots (the V2 and the third quarter of V1) in the genomes of begomoviruses.

INTRODUCTION
The genus Begomovirus (family Geminiviridae) is characterized by dicotyledonous plant-infecting, whitefly-transmitted viruses. Begomoviruses have either monopartite or bipartite genomes that are encapsidated as circular single-stranded DNA (ssDNA) molecules within twin icosahedral particles. During the last two decades, new begomovirus species have emerged worldwide, probably as a consequence of the spread of one or more highly polyphagous biotypes of their insect vector, Bemisia tabaci (Rybicki & Pietersen, 1999). Usually multiple begomovirus species have emerged simultaneously in a given region, with the ensuing species complexes causing diseases in a wide variety of plant species, including many of great agricultural importance.

The rate at which new species are emerging is perhaps best exemplified by the diversity of the almost 700 full begomovirus DNA-A sequences currently deposited in public sequence databases. Given the 89% identity threshold of the International Committee on Taxonomy

Supplementary material is available with the online version of this paper.
of Viruses (ICTV), these genomes represent more than 200 species (Fauquet et al., 2007).

While providing a major contribution to the richness of currently observed begomovirus species diversity, recombination continues both to fuel begomovirus diversification and complicate the classification of new species. The important contribution of recombination to geminivirus evolution is now well established (Umaharan et al., 1998; Padidam et al., 1999) and it is suspected that it is directly responsible for the emergence of many of the most agriculturally damaging begomovirus species complexes (Zhou et al., 1997; Monci et al., 2002; Garcia-Andres et al., 2006). Despite this, very little is actually known either about why recombination seems to contribute to the emergence of species complexes, or how recombinants with enhanced pathogenicity arise and proliferate. Furthermore, both the biochemical processes that determine the kinds of recombinant genomes produced, and the evolutionary processes that determine which of these survive, remain a complete mystery. However, some studies have indicated that recombination hot-spots may exist within begomovirus genomes (Stanley, 1995; Ndunguru et al., 2005; Fauquet et al., 2005; Garcia-Andres et al., 2007). Identifying the locations of any recombination hot- and cold-spots within begomovirus genomes sampled from nature would certainly be a valuable first step towards understanding the underlying processes controlling the generation and spread of recombinants within species complexes.

We decided to quantitatively evaluate the importance of recombination in the genetic diversification of begomoviruses within a newly discovered multipartite begomovirus species complex indigenous to the south-west Indian Ocean (SWIO) islands off the coast of Africa. Despite the pace at which begomovirus diversity has been explored in the past few years, very few full-length African begomovirus DNA-A sequences other than those of the African cassava mosaic disease (CMD) pathosystem are presently available. To increase the richness of the available African begomovirus genome sequence data, we therefore extended previous preliminary surveys of multipartite begomovirus species on the islands of Madagascar, Comoros and Seychelles archipelagos (Lefeuvre et al., 2007; Delatte et al., 2005b). We describe the molecular diversity and taxonomic relationships of 14 SWIO island begomovirus isolates, including seven new species, causing recent plant disease epidemics in the SWIO islands. Importantly, when analysed together with African and Mediterranean begomovirus sequences, we find solid statistical evidence of recombination hot- and cold-spots within the DNA-A components of these viruses. This result may indicate how and why recombination makes such a substantial contribution to begomovirus diversity in general.

METHODS

Sampling and DNA extraction. Tomato (Solanum lycopersicon), tobacco (Nicotiana tabacum) and bean (Phaseolus vulgaris) leaves presenting leaf curling symptoms were collected from individual plants on the islands of the Comoros archipelago (Anjouan, Grande Comore, Mayotte and Mohéli), the Seychelles archipelago (Mahé) and Madagascar (Table 1) and stored dried (Bos, 1977). Total DNA was extracted using a DNeasy Plant miniprep kit (Qiagen) according to the manufacturer’s instructions.

PCR detection. Polymerase chain reaction (PCR) was used to amplify two fragments from the extracted DNA of all samples using two degenerate primer sets: AV494-AC1048 (Wyatt & Brown, 1996), and VD360-CD1266 (Delatte et al., 2005b). PCR reactions were carried out as described in Delatte et al. (2005b). The presence/absence of a DNA-B genome component and DNA-c components were also assessed for each of the isolates using, respectively, the PCR primer sets PBL1v2040-PCRc1 (Rojas et al., 1993) and Beta 1-Beta 2 (Bridgon et al., 2002).

Cloning strategies. Circular viral DNA molecules were amplified using a TempliPhi kit (GE Healthcare) as described by Inoue-Nagata et al. (2004). Full genomes were cloned into the vector pBc-KS in the HindIII restriction site for AM701758, AM701759, AM701766, AM701767 and AM491778 and in the BamHI restriction site for all others. A complete DNA-A-like component for each isolate was sequenced by gene walking using the Macrogen sequencing service (Korea).

Phylogenetic analysis. Full DNA-A-like sequences from 14 isolates (this study) were arranged so that the first nucleotide in the sequence corresponded to the first base (adenine) of virion strand replication (Laufs et al., 1995). Forty-one other full DNA-A and DNA-A-like sequences of related viruses were obtained from public sequence databases using TaxBrowser (http://www.ncbi.nlm.nih.gov/) on May 2006. Multiple sequence alignments were constructed using partial order graphs (POA) (Lee et al., 2002), the CLUSTAL W (Thompson et al., 1994) based subalignment tool available in MEGA 3.1 (Kumar et al., 2004) and manual editing.

The optimal model of sequence evolution defined by ModelTest (Posada, 2006) was used for phylogenetic reconstruction (GTR+1+I). The maximum-likelihood (ML) tree was determined from a preliminary neighbour-joining (NJ) analysis using PAUP* with the heuristic search algorithm. In addition to these analyses, we performed Bayesian phylogenetic reconstruction on the full dataset using MrBayes 3.1.2 (Ronquist & Huelsenbeck, 2003). Four runs with six Markov chains were conducted simultaneously for 1 000 000 generations starting from random initial trees, and sampled every 100 generations. Variation in the ML scores in this sample was examined graphically with Tracer (Rambaut & Drummond, 2004). The trees generated prior to stabilization of ML scores were discarded with the consensus phylogeny and posterior probability of their nodes being determined with a burn-in of 25%. The method of Shimodaira & Hasegawa (1999) implemented in PAUP* was used to test whether the ML scores of the NJ, ML and Bayesian phylogenetic reconstructions fell within the same confidence limits.

Recombination analyses. Detection of potential recombinant sequences, identification of likely parental sequences and localization of possible recombination breakpoints was carried out on a 178-sequence alignment (170 begomovirus, seven curtovirus and one topocovirus sequences) using the RDP (Martin & Rybicki, 2000), GENECOVN (Padidam et al., 1999), BOOTSCAN (Martin et al., 2005a), MAXIMUM CHI SQUARE (Smith, 1992), CHIMAERA (Martin et al., 2005b) and SISTER SCAN (Gibbs et al., 2000) recombination detection methods as implemented in RDP3 (Martin et al., 2005b), available from http://darwin.uvigo.es/rdp/rdp.html (see the RDP project file submitted as supplementary material for full details of program settings). The analysis was performed with default settings for the different
Table 1. Geographical origin and characterization of SWIO begomovirus isolates

<table>
<thead>
<tr>
<th>Region</th>
<th>Island or province/ district</th>
<th>Village</th>
<th>Host plant</th>
<th>Year</th>
<th>Acronym</th>
<th>Closest available virus (%) identity</th>
<th>DNA length</th>
<th>Predicted coding capacity (amino acid)</th>
<th>First description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comoros archipelago</td>
<td>Anjouan</td>
<td>Ouani</td>
<td>Tomato</td>
<td>2004</td>
<td>ToLCAnV-[Anj:Oua3:04]*</td>
<td>ToLCYTV-[Dem] (96 %)</td>
<td>2781 116 258 389 135 134 77†</td>
<td>This study</td>
<td></td>
</tr>
<tr>
<td>Grande Comore</td>
<td>Simboussa</td>
<td>Tobacco</td>
<td>2004</td>
<td>ToLCMVK-[GC:Sim18:04]*</td>
<td>ToLCYTV-[Dem] (82 %)</td>
<td>2774 98† 258 359 135 134 100</td>
<td>This study</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mayotte</td>
<td>Dimadjou</td>
<td>Tobacco</td>
<td>2004</td>
<td>ToLCAnV-[GC:Dim44:04]*</td>
<td>ToLCYTV-[Dem] (83 %)</td>
<td>2755 116 258 358 135 134 85†</td>
<td>This study</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fouboudziouini</td>
<td>Tobacco</td>
<td>2005</td>
<td>TbLCZV-[GC:Fbr95:05]</td>
<td>ToLCYTV-[Dem] (96 %)</td>
<td>2764 116 258 371ll 135 134 85§</td>
<td>This study</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dembeni</td>
<td>Tobacco</td>
<td>2005</td>
<td>TbLCZV-[YT:Dem03]</td>
<td>ToLCYTV-[Dem] (83 %)</td>
<td>2758 133ll 258 358 135 134 100</td>
<td>This study</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kahani</td>
<td>Tobacco</td>
<td>2003</td>
<td>ToLCYTV-[YT:Kah:03]</td>
<td>ToLCMVK-[Dem] (88 %)</td>
<td>2765 116 258 358 135 134 100</td>
<td>Delatte et al. (2005b)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Madagascar</td>
<td>Namakely</td>
<td>Tomato</td>
<td>2001</td>
<td>ToLCAnV-[MG:Nam3:01]</td>
<td>ToLCYTV-[Dem] (86 %)</td>
<td>2769 116 258 359 135 134 85†</td>
<td>This study</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Antsalaka</td>
<td>Tomato</td>
<td>2001</td>
<td>ToLCAnV-[MG:Ant6:01]*</td>
<td>ToLCYTV-[Dem] (86 %)</td>
<td>2754 122 258 359 135 134 85§</td>
<td>This study</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Miandrivazo</td>
<td>Tomato</td>
<td>2001</td>
<td>ToLCYTV-[MG:Mia1:01]*</td>
<td>ToLCYTV-[Dem] (86 %)</td>
<td>2775 116 258 359 135 134 100</td>
<td>This study</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Miandrivazo</td>
<td>Tomato</td>
<td>2001</td>
<td>ToLCYTV-[MG:Mia2:01]*</td>
<td>ToLCYTV-[Dem] (82 %)</td>
<td>2764 116 258 371ll 135 134 85§</td>
<td>This study</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fort Dauphin</td>
<td>Bean</td>
<td>2001</td>
<td>CLCuGV-Be-[MG:For01]</td>
<td>ToLCYTV-[Dem] (86 %)</td>
<td>2754 122 258 362 134 100</td>
<td>Delatte et al. (2005b)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Morondava</td>
<td>Tomato</td>
<td>2001</td>
<td>ToLCMGV-Men-[MG:Mor01]</td>
<td>ToLCYTV-[Dem] (86 %)</td>
<td>2777 116 258 359 135 134 100</td>
<td>Delatte et al. (2005b)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Atsimo</td>
<td>Tomato</td>
<td>2001</td>
<td>ToLCMGV-Ats-[MG:To10]</td>
<td>ToLCYTV-[Dem] (86 %)</td>
<td>2775 116 258 359 135 134 100</td>
<td>Delatte et al. (2005b)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mahé</td>
<td>Tomato</td>
<td>2004</td>
<td>ToLCSCV-[Mah:VE77:04]*</td>
<td>ToLCYTV-[Dem] (81 %)</td>
<td>2742 116 258 375ll 135 134 85§</td>
<td>This study</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*New species proposal.
†ORF containing a premature stop codon.
§ORF containing a frame-shift mutation.
§No predicted ORF identified.
||ORF containing an in-frame ATG codon upstream of the putative initiation codon.
detection methods and a Bonferroni corrected P-value cut-off of 0.05. The breakpoint positions and recombinant sequence(s) inferred for every detected potential recombination event were manually checked and adjusted where necessary using the extensive phylogenetic and recombination signal analysis features available in RDP3. Once a set of unique potential recombination events was identified, we compiled a breakpoint map by plotting the positions of all clearly identifiable breakpoints. A breakpoint density plot was then constructed from this map and the statistical significance of potential breakpoint hot- and cold-spots was tested as described in Heath et al. (2006). Briefly, the statistical analysis used takes the observed distribution of polymorphic sites in an alignment and randomly maps all the observed recombination events to this distribution, such that the real and randomly mapped events all involve exchanges of sequence tracts containing the same numbers of polymorphic sites. Doing this accounts for the fact that uneven distribution of polymorphic sites along the length of an alignment makes the identification of breakpoints in certain alignment regions more difficult than in others. This random mapping process is then repeated 1000 times and the actual distribution of breakpoints is compared to that of the 1000 permuted mappings using two tests. The first is a ‘global’ test which determines whether there are breakpoint clusters in the real distribution with more breakpoints than generally occur in the distributions determined from the permuted datasets. This analysis is highly conservative as it ignores the fact that it will be far harder to detect a genuinely significant breakpoint cluster in regions of conserved sequence than it will be to detect one in regions of more diverse sequence (as mentioned above, breakpoints are most easily and accurately detectable where diversity is high). Therefore, a second, less conservative, ‘local’ test compares corresponding portions of the real and permuted breakpoint distributions and determines whether local regions of the real distribution contain significantly more breakpoints than generally occur in corresponding regions of the permuted datasets. The P-values associated with both the global and local tests are simply the proportions of permuted datasets with greater breakpoint clusters. Whilst we judged P-values \(<0.05\) to be significant for the conservative global test, to guard against false positives, we judged P-values \(<0.01\) as being significant for the less conservative local test.

**Species distinction analysis.** Sequence identity was computed from the precedent multiple sequence alignments without curtovirus and topocuvirus sequences (170 sequences) using the dna.dist function available in the R package, ape (Paradis et al., 2004). We identified genotypes belonging to different species using the ICTV-recommended \(89\%\) complete DNA-A/DNA-A-like sequence identity threshold for species demarcation. To take into account possible influences of discovery order on species number estimates, we repeated the species identification operation 1000 times using the sequences in a random order. The mean and standard deviation of identified species numbers were calculated from the results of these permutations.

**RESULTS**

**Cloning and sequencing**

The complete nucleotide sequences of 14 DNA-A-like components were determined from dried leaf extracts originating from five different SWIO islands (Table 1; Fig. 1). While PCR amplification and cloning of apparently full-length DNA-A-like components was possible from all symptomatic leaf samples, DNA-B and DNA-\(\beta\) specific PCRs yielded no amplification products. This implied that the 14 viruses were most likely all monopartite, as it has been shown previously for four SWIO species with agroinfectious clones (Delatte et al., 2005b). These DNA-A-like sequences were all of typical monopartite begomovirus size, ranging from 2742 to 2781 nt. Most of the sequences had predicted genes typical of monopartite begomoviruses in terms of both size and position. However, for some sequences, deduced protein sequences contained potential translation errors, as indicated in Table 1. All sequences are available in GenBank/EMBL/DDJ/P database under the accession numbers given in Table 1.

**Species distinction**

On the basis of nucleotide identity to their closest known relatives (Table 1), 7 of the 14 new sequences represent new species. Species names for these viruses, based on host plant and region of origin, are proposed in Table 1. The remaining seven sequences share \(>89\%\) identity with DNA-A-like sequences of previously described species such as Cotton leaf curl Gezira virus (CLCuGV), Tomato leaf curl Comoros virus (ToLCKMV), Tomato leaf curl Mayotte virus (ToLCYTV), Tomato leaf curl Madagascar virus.
contains ACMV, (EACMKV). G3 also contains ToLCMGV isolates from the (EACMZV) and East African cassava mosaic Kenya virus East African cassava mosaic Zanzibar virus related species, such as East African cassava mosaic virus isolates fall into phylogroup G2. Phylogroup G3 contains However, none of the currently described indigenous SWIO species, including African ToLCVs from Sudan and Mali. (TYLCV) isolates and closely related tomato infecting virus contains Mediterranean tomato yellow leaf curl virus (ToLCTolV) are new species; Table 1. Phylogenetic analysis Phylogenetic reconstruction was achieved under the sequence evolutionary model GTR +I+G. The SH-test performed on NJ, ML and Bayesian trees concluded that both the ML and Bayesian phylogenetic reconstructions were congruent and had the greatest likelihood. Most of the nodes of the Bayesian phylogenetic tree had probabilities values greater than or equal to 0.95, indicating that branches are relatively stable (Fig. 2). The Bayesian phylogenetic tree clearly indicated that the African, Mediterranean and SWIO sequences separate into four major clades or phylogroups (G1, G2, G3 and G4; Fig. 2). The SWIO isolates are found in three of these phylogroups (G1, G3 and G4). Viruses widely sampled from various host species (chayote, cotton, hollyhock, pepper, tobacco and tomato) throughout Africa are found in G1, which also contains four SWIO isolates [of which, Tomato leaf curl Diana virus (ToLCDiaV) and Tomato leaf curl Toliara virus (ToLCtoIv) are new species; Table 1]. Phylogroup G2 contains Mediterranean tomato yellow leaf curl virus (TYLCV) isolates and closely related tomato infecting virus species, including African ToLCVs from Sudan and Mali. However, none of the currently described indigenous SWIO isolates fall into phylogroup G2. Phylogroup G3 contains East African cassava mosaic virus (EACMV), East African cassava mosaic Zanzibar virus (EACMZV) and East African cassava mosaic Kenya virus (EACMKV). G3 also contains ToLCMGV isolates from the west coast of Madagascar. Finally, the forth phylogroup, G4, contains ACMV, Tomato leaf curl Uganda virus-[Iganga] (ToLCUGV-[Ig]) and twelve SWIO begomovirus isolates including five new species: Tomato leaf curl Moheli virus (ToLCMohV), Tobacco leaf curl Comoros virus (TbLCKMV), Tomato leaf curl Seychelles virus (ToLSCV), Tomato leaf curl Anjouan virus (ToLCanjV) and Tomato leaf curl Antsiranana virus (ToLCAntV). Importantly, G3 and G4 contain both multipartite and bipartite begomoviruses.

Analysis of recombination We analysed evidence of recombination in a 178-sequence alignment containing 170 full-length SWIO, African and Mediterranean begomovirus DNA-A and DNA-A-like sequences, and eight curtovirus and one topocuvirus full genome sequences. It was apparent from this analysis that collectively the SWIO isolates bear detectable evidence of at least 22 past recombination events (Fig. 3). Only CLCuGV-Be[An:For:01] was not detectably recombinant. Among the recombination events that were detected, many were between different species: the TbLCKMV-[GC:Fou99:05] and TbLCKMV-[GC:Sim18:04] isolates have apparently obtained almost their entire CP ORF from a virus resembling TbLCZV-[ZW] (event ‘p’ in Fig. 3), whereas the rest of their genome resembles that of the tomato infecting virus ToLCAnjV-[MG:Mia1:01] (AM701767). Another very striking recombination event was detected in the ToLCSCV-[Mah:VE77:04] sequence from the Seychelles archipelago (event ‘v’ in Fig. 3). We were surprised to find that part of the Rep ORF of this virus was apparently derived from a divergent begomovirus resembling Sweet potato leaf curl virus (SPLCGV). However, upon closer analysis, it is probably more feasible that both SPLCGV and ToLCSCV-[Mah:VE77:04] have obtained large portions of their C1 ORFs from a curtovirus-like source.

Eighteen out of a total of 22 unique events detected in the SWIO sequences were within the rep gene, indicating that the rep gene in general, and the sequences encoding the Rep N-terminal region in particular, might be a recombination hot-spot. To test this hypothesis, we plotted all unambiguously detectable breakpoint positions on a breakpoint density map and used a permutation test to determine whether the breakpoint distribution was significantly non-random (Fig. 4). This analysis revealed one large ‘globally’ significant recombination hot-spot (global P-values <0.05 across its length) and two smaller ‘locally’ significant hot-spots (local P-values <0.01). Whereas the large global hot-spot encompasses almost the entire intergenic region (IR) between the C1 start codon and approximately 50 nucleotides 5' of the V2 ORF start codon, the locally significant hot-spots occur at the V1–C3 interface and in the centre of the C1 ORF. In addition to these hot-spots, the analysis also revealed two locally significant recombination cold-spots (local P-value <0.01). These occurred in the V2 ORF and in the third quarter of the V1 ORF (Table 2).

DISCUSSION

We have demonstrated that the SWIO islands harbour an extraordinary diverse begomovirus population. On the
Fig. 2. Phylogenetic tree indicating the relationships between the DNA sequences of SWIO begomoviruses and those of a representative sampling of publicly available African and Mediterranean begomovirus sequences. Major clades or phylogroups are labelled G1 through G4. The tree was constructed using MrBayes and rooted using ToLCV-[AU] and ToLCBV as outliers. Numbers associated with nodes indicate the posterior probability for those nodes. Whereas horizontal bars represent genetic distances as indicated by the scale bar, vertical distances are arbitrary. SWIO sequences are in bold and bipartite begomoviruses are underlined. Four phylogenetic groups (G1 to G4) have been defined and are represented by vertical lines.
<table>
<thead>
<tr>
<th>Event</th>
<th>Region</th>
<th>Minor</th>
<th>Major</th>
<th>Detected by</th>
<th>$P$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>2023-2329</td>
<td>TolCNIV-[MG:Ant6:01] AM701766 &amp; TolCNIV-[MG:Ant6:01] AM701767</td>
<td>TolCNIV-[MG:Ant6:01] AM701766 &amp; TolCNIV-[MG:Ant6:01] AM701767</td>
<td>RGBMCs</td>
<td>7.0.10 \textsuperscript{7}</td>
</tr>
</tbody>
</table>
basis of the 14 new complete DNA-A sequences and in accordance with the ICTV guidelines, we describe seven new island begomovirus species. Taken together, ten of the 18 complete DNA-A-like sequences so far determined for SWIO begomoviruses (this study and Delatte et al., 2005b) represent new species. Interestingly, the ten viral species described in the SWIO islands are distributed amongst three of the four major phylogenetic groups identified within the African/Mediterranean begomovirus cluster. The presence of SWIO monopartite begomoviruses in two

of the groups containing the bipartite cassava mosaic begomoviruses (G3 and G4) supports the hypothesis that either there have been multiple DNA-B component losses to produce the three different African monopartite virus lineages, or there have been multiple acquisitions of DNA-B components to produce the bipartite virus lineages (Saunders et al., 2002; Mansoor et al., 2003).

Our attempts at provisional classification of the novel virus genotypes described in this study led us to examine begomovirus species demarcation criteria. Given the prevailing demarcation criteria, we determined that, amongst the 170 sequences examined, the 89 % ICTV begomovirus species demarcation criterion implies that only 49 of these should be classified as species. This analysis revealed something quite interesting about the begomovirus genomes examined. Whereas the genomes contained approximately 43.5 ± 1.1 distinct rep genes, they contained far fewer different kinds of other genes (ranging from 25 ± 1.3 distinct V2 ORFs to 32 ± 1.1 distinct cp genes). There is the equivalent of 36 % more rep genes in circulation than all other SWIO/African/Mediterranean begomovirus genes. Our results clearly indicate that recombination is almost certainly the driving force behind this apparent proliferation of rep genes. This simple result clearly illustrates how recombination confounds the definition of useful taxonomic criteria (Seal et al., 2006; Fauquet et al., 2003).

The phylogenetic analysis performed in this study clearly demonstrates that the breadth of begomovirus diversity found on the SWIO islands is qualitatively similar to that

### Table 2. Recombination hot-spots and cold-spots

<table>
<thead>
<tr>
<th>Type</th>
<th>Position*</th>
<th>Region</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hot-spot</td>
<td>1–50</td>
<td>IR</td>
<td>Globally</td>
</tr>
<tr>
<td></td>
<td>920–1030</td>
<td>V1–C3 interface</td>
<td>Locally</td>
</tr>
<tr>
<td></td>
<td>2070–2210</td>
<td>C1 centre</td>
<td>Locally</td>
</tr>
<tr>
<td>Cold-spot</td>
<td>120–190</td>
<td>V2</td>
<td>Locally</td>
</tr>
<tr>
<td></td>
<td>510–860</td>
<td>V1 third quarter</td>
<td>Locally</td>
</tr>
</tbody>
</table>

*Relative to ToLCKMV-[YT:Dem:03] (GenBank accession no. AJ865341).
identifiable across the entire African continent. Also, besides six isolates classified as belonging to the G1 and G3 groups, the island isolates are all most closely related to one another and only share a distant common ancestor with the mainland viruses. This probably indicates that the SWIO islands have, with the exception of infrequent transmission events from mainland Africa, been epidemiologically isolated for a long time.

Phylogeographically the results are also intriguing: for the G4 group, there is a well supported cluster of SWIO isolates with ACMV. There is evidence here that either (i) ACMV, a lonely outlier amongst the other African viruses, originated on the SWIO islands, or (ii) the SWIO isolates with ACMV. There is evidence here that either (i) ACMV, a group isolated for a long time.

Our recombination analysis clearly indicates the presence of breakpoint hot- and cold-spots within SWIO/African/Mediterranean begomovirus genomes. This indicates either that DNA breakage and repair do not occur randomly in begomoviruses or that, if breakpoints do occur randomly, selection has preferentially culled recombinants with breakpoints in certain positions while permitting the survival of recombinants with breakpoints in other positions. That all recombinants are not created equal has been clearly demonstrated with laboratory constructed geminivirus recombinants (Liu et al., 2001; Martin & Rybicki, 2002) and one would expect original (and possibly still the natural) hosts of these viruses to be indigenous uncultivated plants. Further studies should aim to characterize begomovirus diversity in these hosts.

Importantly, experimental analyses of recombination in geminiviruses (Schnippenkoetter et al., 2001; Stenger et al., 1991; Garcia-Andres et al., 2007) and the replicational release mechanisms put into practice during agroinoculation of geminiviruses, have indicated that the origin of virion strand replication is a biochemically predisposed recombination hot-spot. While there is a clear breakpoint distribution peak detected at the virion strand ori, the highest breakpoint distribution peak is 5′ of the ori, close to the rep start codon. This region corresponds to the most variable region of begomovirus genomes. It is probable that at least part of the reason why so many breakpoints are detected here is that this is the genome region where breakpoints are easiest to detect. Nevertheless, the statistical test used to detect hot spots takes this increased variability into account and has still identified that there are an improbably large number of breakpoints in this region. We propose first, that the IR-wide breakpoint hot-spot is a consequence of recombinants with breakpoints outside of genes generally being fitter than those with breakpoints within genes. This possibility is supported by the fact that the V1–C3 interface, the only other genome region where breakpoints are possible outside of genes, is also a recombination hot-spot.

Importantly, there exists direct experimental support for our observation that the V1–C3 interface is a recombination hot-spot because recombination at this point does not incur a significant fitness cost. In experimental recombination in controlled mixed TYLCSV and TYLCV-Mld infection, the most prevalent (and hence probably the most fit) emergent recombinant had one breakpoint within 100 nucleotides of the V1–C3 ORF interface and another at precisely the virion strand ori (Garcia Andres et al., 2007). That this particular recombinant genotype is highly fit is further evidenced by its close resemblance to the widespread natural TYLCSV–TYLCV-Mld recombinant, Tomato yellow leaf curl Malaga virus (Monci et al., 2002). The problem remains, however, to explain the recombination hot-spot in the middle of the rep gene. Our second proposal is therefore that the N-terminal portion of Rep and any protein expressed from the C4 ORF are exceptionally tolerant of recombination, with the most tolerable breakpoint positions (i.e. those that disrupt Rep folding the least) occurring near the centre of the gene around the recombination hot-spot.

The presence of recombination cold-spots within the V2 ORF and the third quarter of the V1 ORF is consistent with our first proposal that recombination breakpoints within coding regions are generally more damaging than those outside of coding regions. However, the fact that the detectable breakpoint cold-spots are within the virion sense ORFs, whereas the greatest number of breakpoints are
within the complementary sense ORFs, leads us to a third proposal: the uneven distribution of recombination breakpoints is possibly due, at least in part, to clashes between virion strand replication and gene transcription. Whereas replication and virion strand transcription proceed in the same direction and are therefore unlikely to interfere with one another, transcription of the complementary strand ORFs tends to disrupt replication forks moving in the opposite direction. Analysis of replicating begomoviral DNA intermediates has revealed a wide distribution of so-called heterogeneous length linear dsDNA forms (hDNA), possibly created during such clashes. The ends of these hDNA molecules tend to map most frequently to the V-ori and either the C2/C3 transcription promoter near the hot-spot we detected in the centre of rep, or the C2/C3 terminator near the hot-spot we detected at the V1–C3 ORF interface (Jeske et al., 2001). Completion of replication from displaced, partially replicated virion strands would then proceed via the recombination-dependent replication pathway (Preiss & Jeske, 2003), which in the presence of potential template DNAs with different sequences could result in detectable recombination events. Completion of replication would result in a recombinant virion strand with one breakpoint at the point where replication was initially disrupted and the other at the virion sense ori where replication was completed.

Novel environments, such as the new host species offered to begomoviruses by invasive polyphagous vector biotypes, are possibly the defining force driving begomovirus evolution worldwide. For example, introduction into Reunion of the polyphagous B. tabaci biotype B is believed to be responsible for severe TYLC disease epidemics on the island in the late 1990s (Peterschmitt et al., 1999; Delatte et al., 2005a). Spread of this biotype to other SWIO islands may (i) facilitate host switching into cultivated crops of uncharacterized begomoviruses that currently only infect weeds and (ii) induce an overlap of exotic TYLCV and indigenous begomovirus distributions. Given the propensity of begomoviruses to recombine, emergence of new recombinants with increased virulence and/or modified host ranges are to be expected. An emergent TYLCV– Tomato yellow leaf curl Sardinia virus (TYLCSV) recombinant lineage in Spain (Monci et al., 2002; Garcia-Andres et al., 2006) demonstrates that the probability of such an occurrence is high, especially as the genetic distance between TYLCV and the SWIO indigenous ToLCVs is similar to the distance between TYLCV and TYLCSV.

By highlighting the extraordinary diversity of begomoviruses on the SWIO islands, we have provided a detailed description of their phylogenetic and recombinant histories. The phylogenetic association between the monopartite SWIO isolates and both monopartite and bipartite mainland African isolates indicate that they are probably indigenous to the islands. The large number of unique recombination events that we have detected amongst the SWIO isolates and their nearest mainland relatives reiterates the pivotal role of this process in begomovirus evolution. It is, however, apparent from our breakpoint distribution analysis that purifying selection and/or varying biochemical predispositions to recombination in different parts of begomovirus genomes place substantial constraints on the degree of evolutionary innovation that is possible by recombination.

ACKNOWLEDGEMENTS

This work was funded by the Conseil Régional de la Réunion, the Ministère de l’Outre-Mer, CIRAD and the MRES. D. P. M. is funded by the Harry Oppenheimer Foundation, a Sydney Brenner Fellowship and the South African Bioinformatics Network. A.V. is supported by the Carnegie Corporation of New York.

REFERENCES


