Journal of General Virology (2007), **88**, 3458–3468 **DOI** 10.1099/vir.0.83252-0

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. Reynaud[1] and J.-M. Lett[1]

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

[1]CIRAD, UMR 53 PVBMT CIRAD-Université de la Réunion, Pôle de Protection des Plantes, 7 Chemin de l'IRAT, 97410 Saint Pierre, La Réunion, France

[2]Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, Observatory 7925, South Africa

[3]Electron Microscopy Unit, University of Cape Town, Rondebosch 7701, South Africa

[4]Museum National d'Histoire Naturelle, Dept RDDM, USM 501, CNRS UMR 5166, Evolution des Régulations Endocriniennes, 57 rue Cuvier, CP 32, 75005 Paris, France

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.

Received 20 June 2007
Accepted 9 August 2007

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 monopartite 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 monopartite 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 tabaci*) 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 Moheli), 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-β molecules 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 (Briddon *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 *Hin*dIII restriction site for AM701758, AM701759, AM701766, AM701767 and AM491778 and in the *Bam*HI 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+I+G). 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 topocuvirus sequences) using the RDP (Martin & Rybicki, 2000), GENECONV (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

ToLCAnjV-[Anj:Oua3:04], Tomato leaf curl Anjouan virus - [Anjouan:Ouani3:2004] (GenBank accession no. AM701758); ToLCKMJV-[Anj:Bam5:04], Tomato leaf curl Comoros virus - [Anjouan:Bambas5:2004] (AM701759); TbLCKMV-[GC:Sim18:04], Tobacco leaf curl Comoros virus - [Grande Comore:Simboussa18:2004] (AM701760); ToLCAntV-[GC:Dim44:04], Tomato leaf curl Antsiranana virus - [Grande Comore:Dimadjou44:2004] (AM701761); TbLCZV-[GC:Fbz95:05], Tobacco leaf curl Zimbabwe virus - [Grande Comore:Foumboudziouni95:2005] (AM701756); TbLCKMV-[GC:Fou99:05], Tobacco leaf curl Comoros virus - [Grande Comore:Foubouni99:2005] (AM701762); ToLCKMV-[YT:Dem:03], Tomato leaf curl Comoros virus - [Mayotte:Dembeni:2003] (AJ865341); ToLCYTV-[YT:Kah:03], Tomato leaf curl Mayotte virus - [Mayotte:Kahani:2003] (AJ865340); ToLCMohV-[KM:Fom163:05], Tomato leaf curl Moheli virus - [Comoros:Fomboni163:2005] (AM701763); ToLCAntV-[MG:Nam3:01], Tomato leaf curl Antsiranana virus - [Madagascar:Namakely3:2001] (AM701764); ToLCDiaV-[MG:Nam5:01], Tomato leaf curl Diana virus - [Madagascar:Namakely5:2001] (AM701765); ToLCAntV-[MG:Ant6:01], Tomato leaf curl Antsiranana virus - [Madagascar:Antsalaka6:2001] (AM701766); ToLCAntV-[MG:Mia1:01], Tomato leaf curl Antsiranana virus - [Madagascar:Miandrivazo1:2001] (AM701767); ToLCTolV-[MG:Mia2:01], Tomato leaf curl Toliara virus - [Madagascar:Miandrivazo2:2001] (AM701768); CLCuGV-Be[MG:For:01], Cotton leaf curl Gezira virus - Bean[Madagascar:Fort Dauphin:2001] (AM701757); ToLCMGV-Men[MG:Mor:01], Tomato leaf curl Madagascar virus - Menabe[Madagascar:Morondava:2001] (AJ865338); ToLCMGV-Ats[MG:Tol:01], Tomato leaf curl Madagascar virus - Atsimo[Madagascar:Toliary:2001] (AJ865339); ToLCSCV-[Mah:VE77:04], Tomato leaf curl Seychelles virus - [Mahé:Val d'Endor77:2004] (AM491778).

Region	Island or province/ district	Village	Host plant	Year	Acronym	Closest available virus (% identity)	DNA length	Predicted coding capacity (amino acid)						First description
								V2	V1	C1	C2	C3	C4	
Comoros archipelago	Anjouan	Ouani	Tomato	2004	ToLCAnjV-[Anj:Oua3:04]*	ToLCYTV-[Dem] (96 %)	2781	116	258	389	135	134	77†	This study
		Bambas	Tomato	2004	ToLCKMJV-[Anj:Bam5:04]	ToLCYTV-[Dem] (82 %)	2774	98‡	258	359	135	134	100	This study
	Grande Comore	Simboussa	Tobacco	2004	TbLCKMV-[GC:Sim18:04]*	ToLCYTV-[Dem] (83 %)	2755	116	258	358	135	134	100	This study
		Dimadjou	Tomato	2004	ToLCAntV-[GC:Dim44:04]*	ToLCMGV-[Mor] (86 %)	2772	98‡	258	215†	135	109‡	§	This study
		Foumboudziouni	Tobacco	2005	TbLCZV-[GC:Fbz95:05]	TbLCZV-[ZW] (96 %)	2764	116	258	371‖	135	134	85†	This study
		Foubouni	Tobacco	2005	TbLCKMV-[GC:Fou99:05]*	ToLCYTV-[Dem] (83 %)	2758	133‖	258	358	135	134	100	This study
	Mayotte	Dembeni	Tomato	2003	ToLCKMV-[YT:Dem:03]	ToLCYTV-[Dem] (88 %)	2765	116	258	358	135	134	100	Delatte *et al.* (2005b)
		Kahani	Tomato	2003	ToLCYTV-[YT:Kah:03]	ToLCKMV-[Kah] (88 %)	2768	116	258	379	135	134	143	Delatte *et al.* (2005b)
	Mohéli	Fomboni	Tomato	2005	ToLCMohV-[KM:Fom163:05]*	ToLCYTV-[Dem] (88 %)	2756	118	258	235†	135	134	100	This study
Madagascar	Antsiranana/Diana	Namakely	Tomato	2001	ToLCAntV-[MG:Nam3:01]*	ToLCYTV-[Dem] (86 %)	2769	116	258	359	135	134	100	This study
	Antsiranana/Diana	Namakely	Tomato	2001	ToLCDiaV-[MG:Nam5:01]*	ClCuGV-[Sha] (82 %)	2754	122	258	359	135	134	85†	This study
	Antsiranana/Diana	Antsalaka	Tomato	2001	ToLCAntV-[MG:Ant6:01]*	ToLCUGV-[Iga] (86 %)	2775	116	258	402‖	135	134	100	This study
	Toliara/Menabe	Miandrivazo	Tomato	2001	ToLCAntV-[MG:Mia1:01]*	ToLCUGV-[Iga] (86 %)	2774	116	258	358	135	134	100	This study
	Toliara/Menabe	Miandrivazo	Tomato	2001	ToLCTolV-[MG:Mia2:01]*	ToLCTZV-[Aru] (83 %)	2764	116	258	376	135	134	85†	This study
	Toliara/Anosy	Fort Dauphin	Bean	2001	CLCuGV-Be[MG:For:01]	ClCuGV-[Ok:Sha] (89 %)	2754	122	258	362	134	133	100	This study
	Toliara/Menabe	Morondava	Tomato	2001	ToLCMGV-Men[MG:Mor:01]	ToLCMGV-[Tol] (94 %)	2777	116	258	359	135	134	100	Delatte *et al.* (2005b)
	Toliara/Atsimo	Toliary	Tomato	2001	ToLCMGV-Ats[MG:Tol:01]	ToLCMGV-[Mor] (94 %)	2775	116	258	359	135	134	100	Delatte *et al.* (2005b)
Seychelles	Mahé	Val d'Endor	Tomato	2004	ToLCSCV-[Mah:VE77:04]*	ToLCYTV-[Dem] (81 %)	2742	116	258	375‖	183	§	85†	This study

*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-β 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/DDJP 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*

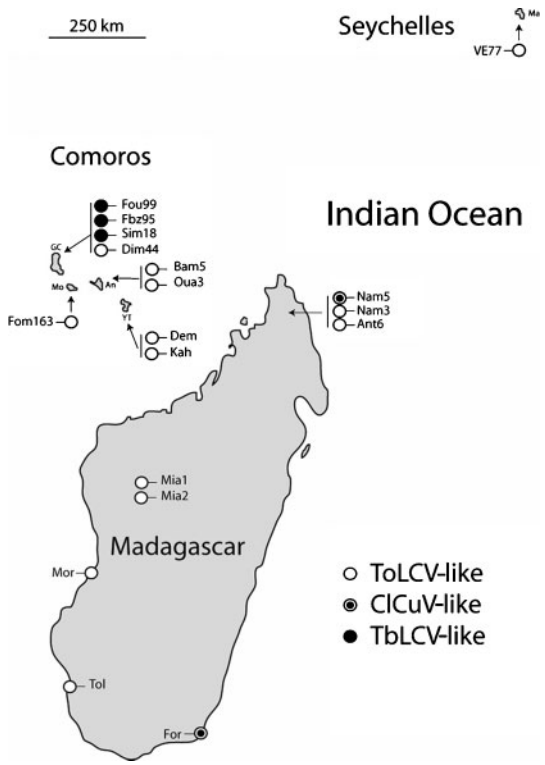


Fig. 1. Map of south-west Indian Ocean islands showing the distribution of begomovirus isolates examined in this study. Colours are used to designate related species. Isolate acronyms: Ant, Antsiranana; Bam, Bambas; Dem, Dembeni; Dim, Dimadjou; Fbz, Foumboudziouni; Fom, Fomboni; Fou, Foubouni; For, Fort Dauphin; Kah, Kahani; Mia, Miandrivazo; Mor, Morondava; Nam, Namakely; Oua, Ouani; Tol, Toliary. Islands acronyms: An, Anjouan; GC, Grande Comore; Ma, Mahé; Mo, Moheli; YT, Mayotte.

(ToLCMGV) and *Tobacco leaf curl Zimbabwe virus* (TbLCZV).

We assembled a dataset of 170 African, Mediterranean and SWIO begomovirus DNA-A-like sequences containing 51 ICTV-designated species and the seven tentative new species described here. We applied our species-counting algorithm test to this dataset and determined that, depending on the discovery order, 49 species ($\pm$1.3) should be defined as such. While this number is not substantially different from 51 classified by the ICTV, our dataset contains an additional seven sequences that are almost certainly legitimate novel species. We further attempted to determine which parts of the virus genomes mostly contribute to pairwise distance scores that are currently the primary taxonomic measure used for species demarcation. Using alignments of individual ORFs indicated that 32$\pm$1.1 (V1 ORF), 25$\pm$1.3 (V2 ORF), 43.5$\pm$1.1 (C1 ORF), 29.5$\pm$1.3 (C2 ORF), 29.5$\pm$1.1 (C3 ORF) and 43.5$\pm$1.3 (C4 ORF) groups of sequences might be defined if we consider only these subgenomes and a 89% identity threshold. This clearly indicated that sequences of the C1 and C4 ORFs are the primary source of taxonomic signal in begomoviruses.

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* (ToLCTolV) 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) and other closely related species, such as *South African cassava mosaic virus* (ACMV), *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-[Iga]) 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 monopartite 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 ToLCAntV-[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 extraordinarily diverse begomovirus population. On the

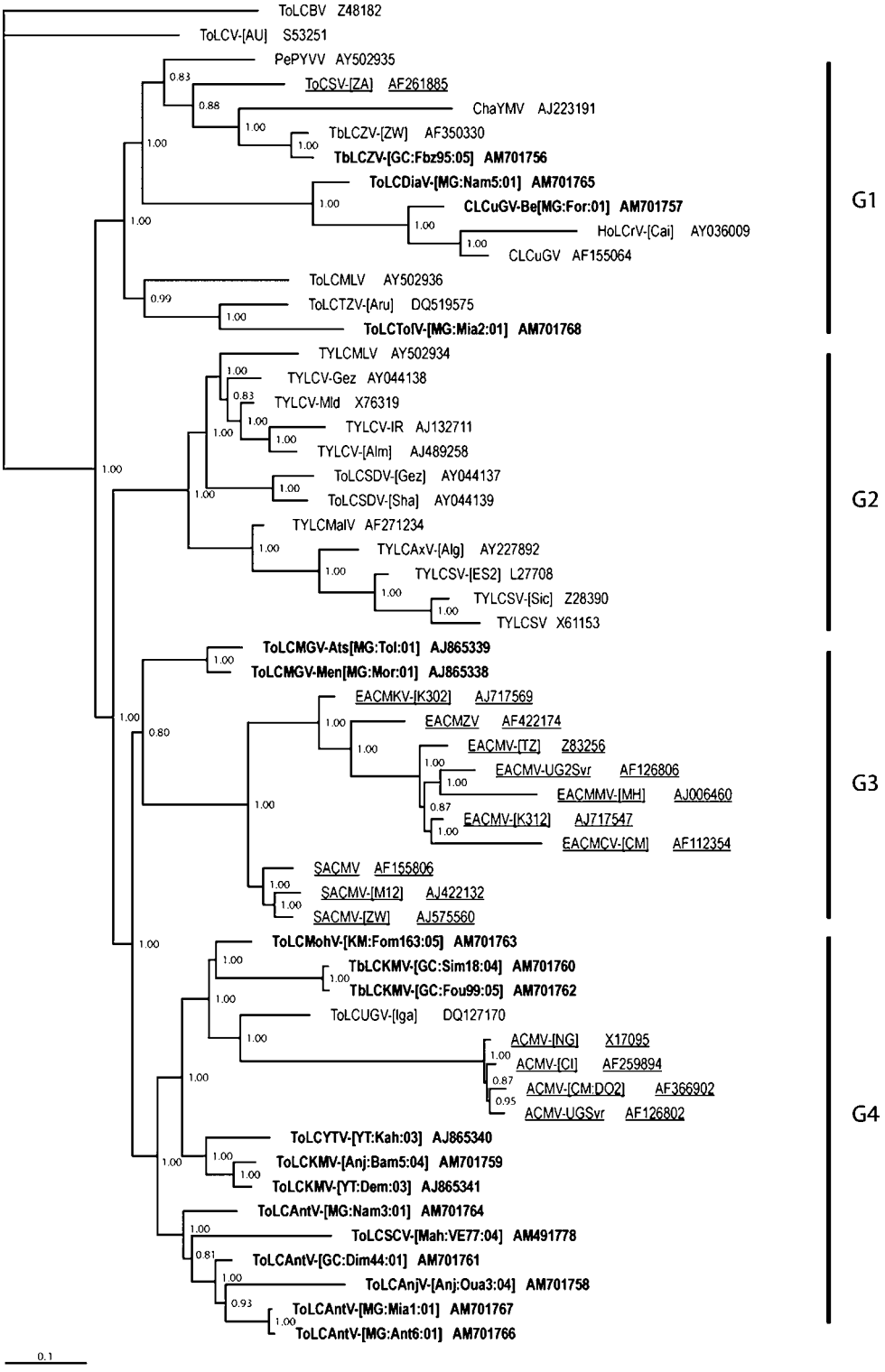
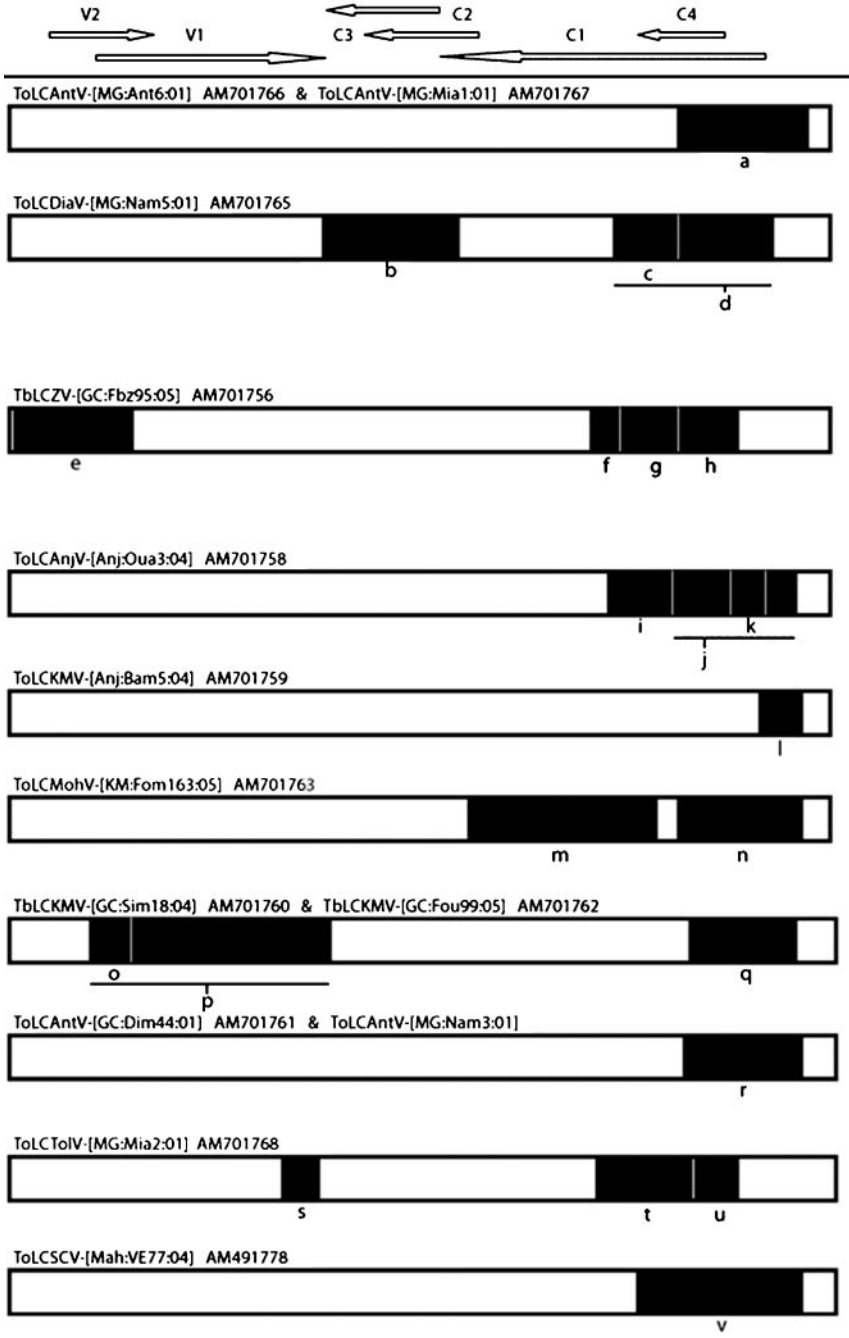


Fig. 2. Phylogenetic tree indicating the relationships between the DNA sequences of SWIO begomoviruses and those of a representative sampling of publicly available African and Mediterranean begomovirus sequences. Major clades or phylogroups are labelled G1 through G4. The tree was constructed using MrBayes 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.



Event	Region	Parental sequences: Minor	Parental sequences: Major	Detected by	*P*-value
a	2293-2721	**ToLCMohV-[KM:Fom163:05] AM701763**	**ToLCAntV-[GC:Dim44:01] AM701761**	RGBMCS	5,6.10 $^{-31}$
b	1080-1528	**ToLCAntV-[MG:Mia1:01] AM701767**	**CLCuGV-Be[MG:For:01] AM701757**	RGBMCS	7,8.10 $^{-31}$
c	2081-2598	BYVMV-[301] AJ002453	Unknown	RGBMC	8,9.10 $^{-8}$
d	2085-2285	AF350330	**CLCuGV-Be[MG:For:01] AM701757**	RGBMCS	1,8.10 ^{18}
e	16-401	PepYVV AY502935	Unknown	RGBMC	4,8.10 $^{-5}$
f	2995-2083	AEV AJ437618	TbLCZV-[ZW] AF350330	RGBm	2,9.10 $^{-13}$
g	2111-2282	Unknown	BYVMV-[301] AJ002453	RGBMC	8,9.10 $^{-8}$
h	2321-2484	CLCuMV-[26] AJ002458	EACMV-[TZ10] AY795988	rmS	1,8.10 $^{-4}$
i	2065-2269	Unknown	**ToLCAntV-[MG:Mia1:01] AM701767**	RGBMC	5,8.10 $^{-30}$
j	2274-2670	Unknown	**ToLCAntV-[GC:Dim44:01] AM701761**	RGBMCS	8,9.10 $^{-27}$
k	2488-2590	Unknown	TYLCSV X61153	RGc	3,6.10 $^{-5}$
l	2578-2698	EACMV-[TZ] Z83256	**ToLCYTV-[YT:Kah:03] AJ865340**	RGBMCS	7,0.10 $^{-21}$
m	1586-2197	**ToLCAntV-[GC:Dim44:01] AM701761**	**ToLCKMV-[Anj:Bam5:04] AM701759**	RGBMCS	4,3.10 $^{-12}$
n	2303-2695	**TbLCKMV-[GC:Fou99:05] AM701762**	**ToLCYTV-[YT:Kah:03] AJ865340**	RGBMC	3,0.10 ^{41}
o	287-403	PePYVV AY502935	SiLCV AM050730	RGBMCS	5,0.10 ^{48}
p	286-1082	**TbLCZV-[GC:Fbz95:05] AM701756**	**ToLCAntV-[MG:Mia1:01] AM701767**	RGBMCS	4,8.10 $^{-5}$
q	2347-2678	ACMV-[KE] J02057	**ToLCYTV-[YT:Kah:03] AJ865340**	RGBMCS	9,2.10 $^{-8}$
r	2326-2709	SiLCV AM050730	EACMV-KE2[K48] AJ717542	RGBmC	1,2.10 $^{-9}$
s	936-1040	**ToLCKMV-[Anj:Bam5:04] AM701759**	Unknown	rgB	2,4.10 $^{-3}$
t	2023-2329	**TbLCZV-[GC:Fbz95:05] AM701756**	EACMV-[K6] AJ717546	RGBMC	7,0.10 $^{-26}$
u	2364-2485	CLCuMV-[26] AJ002458	EACMV-[TZ10] AY795988	rbS	1,8.10 $^{-4}$
v	2149-2703	SPLCGV-[16] AF326775	**ToLCAntV-[MG:Mia1:01] AM701767**	RGBMC	1,0.10 ^{65}

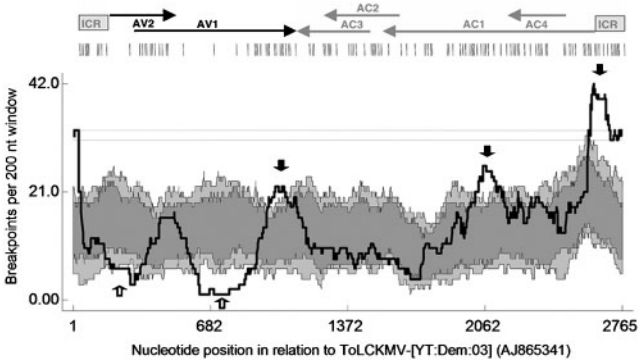


Fig. 4. The distribution of recombination breakpoints within SWIO, African and Mediterranean begomovirus full genome sequences. All detectable breakpoint positions are indicated by small vertical lines at the top of the graph. A 200 nucleotide window was moved along the alignment one nucleotide at a time and the number of breakpoints detected within the window region was counted and plotted (solid line). The upper and lower broken lines, respectively, indicate 99 and 95 % confidence thresholds for globally significant breakpoint clusters. Light and dark grey areas, respectively, indicate local 99 and 95 % breakpoint clustering thresholds taking into account local regional differences in sequence diversity that influence the ability of different methods to detect recombination breakpoints. Vertical black arrows indicate recombination hot-spots, while vertical white arrows represent recombination cold-spots. Horizontal black arrows represent virion strand ORFs (V1 and V2) and horizontal grey arrows represent complementary strand ORFs (C1, C2, C3 and C4) and the boxes represent the intergenic common region (ICR). The virion strand *ori* is at nucleotide position 1.

Table 2. Recombination hot-spots and cold-spots

Type	Position*	Region	Significance
Hot-spot	1–50	IR	Globally
	920–1030	V1–C3 interface	Locally
	2070–2210	C1 centre	Locally
Cold-spot	120–190	V2	Locally
	510–860	V1 third quarter	Locally

*Relative to ToLCKMV-[YT:Dem:03] (GenBank accession no. AJ865341).

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

Fig. 3. Recombinant regions detected within SWIO virus sequences. The genome at the top of the figure corresponds to the schematic representation of sequences below. Region coordinates are nucleotide positions of detected recombination breakpoints in the multiple sequence alignment used to detect recombination. Wherever possible, parental sequences are identified. 'Major' and 'Minor' parents are sequences that were used, along with the indicated recombinant sequence, to identify recombination. Whereas for each identified event the minor parent is apparently the contributor of the sequence within the indicated region, the major parent is the apparent contributor of the rest of the sequence. Note that the identified 'parental sequences' are not the actual parents but are simply those sequences most similar to the actual parents in the dataset analysed. Recombinant regions and parental viruses were identified using the RDP (R), GENECONV (G), BOOTSCAN (B), MAXIMUM CHI SQUARE (M), CHIMAERA (C) and SISTER SCAN (S) methods. The reported *P*-value is for the method in bold type and is the best *P*-value calculated for the region in question. Whereas upper-case letters imply that a method detected recombination with a multiple comparison corrected $P<0.01$, lower-case letters imply that the method detected recombination with a multiple comparison corrected $P<0.05$ but $\geqslant 0.01$. New virus isolates presented in this study are in bold.

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 are an extant and thriving population of an ancestral lineage that, besides ACMV and ToLCUGV-[Iga], has largely disappeared on the African mainland. As with cassava, the plants sampled in this study are exotic introduced species. 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.

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 that many, if not most, natural recombinants would experience serious fitness deficits. There are in fact two well-supported explanations as to why recombinants are generally less fit than their parents. First, protein engineers have discovered that hybrid genes with bits of sequence from distantly related sources tend to encode proteins that do not fold properly – probably due to disruptions of co-evolved amino acid contacts within their structures (Voigt *et al.*, 2002; Saraf & Maranas, 2003). Second, when genes are transferred wholesale into distantly related genetic backgrounds, they appear only to function well either when they do not interact with a lot of other genes, or when they are co-transferred with the other genes with which they do interact (Jain *et al.*, 1999; Martin *et al.*, 2005c; Escriu *et al.*, 2007). It is therefore likely that, whereas the recombination hot-spots we have detected represent genomic regions where breakpoints are both biochemically permissive and highly 'survivable', the cold-spots represent regions where breakpoints are either particularly deleterious or are biochemically very unlikely to occur.

Although we detected a large number of unique recombination breakpoints (i.e. breakpoints that occurred during different recombination events) across the entire 3′ portion of *rep* spanning the C4 ORF, this region is bounded by two recombination hot-spots. This pattern of recombination is almost certainly due to both a biochemical predisposition to recombination in these sequences, and a high tolerance for recombination in the proteins encoded in this region. 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 Carnegic Corporation of New York.

REFERENCES

Bos, L. (1977). Persistence of infectivity of three viruses in plant material dried over $CaCl_2$ and stored under different conditions. *Eur J Plant Pathol* **83**, 217–220.

Briddon, R. W., Bull, S. E., Mansoor, S., Amin, I. & Markham, P. G. (2002). Universal primers for the PCR-mediated amplification of DNA beta: a molecule associated with some monopartite begomoviruses. *Mol Biotechnol* **20**, 315–318.

Delatte, H., Holota, H., Naze, F., Peterschmitt, M., Reynaud, B. & Lett, J. M. (2005a). The presence of both recombinant and nonrecombinant strains of Tomato yellow leaf curl virus on tomato in Reunion Island. *Plant Pathol* **54**, 262.

Delatte, H., Martin, D. P., Naze, F., Golbach, R. W., Reynaud, B., Peterschmitt, M. & Lett, J. M. (2005b). South West Indian Ocean islands tomato begomovirus populations represent a new major monopartite begomovirus group. *J Gen Virol* **86**, 1533–1542.

Escriu, F., Fraile, A. & Garcia-Arenal, F. (2007). Constraints to genetic exchange support gene coadaptation in a tripartite RNA virus. *PLoS Pathog* **3**, e8.

Fauquet, C. M., Bisaro, D. M., Briddon, R. W., Brown, J. K., Harrison, B. D., Rybicki, E. P., Stenger, D. C. & Stanley, J. (2003). Revision of taxonomic criteria for species demarcation in the family *Geminiviridae*, and an updated list of begomovirus species. *Arch Virol* **148**, 405–421.

Fauquet, C. M., Sawyer, S., Idris, A. M. & Brown, J. K. (2005). Sequence analysis and classification of apparent recombinant begomoviruses infecting tomato in the Nile and Mediterranean Basins. *Phytopathology* **95**, 549–555.

Fauquet, C. M., Briddon, R. W., Brown, J. K., Moriones, E., Stanley, J., Zerbini, M. & Zhou, X. (2007). Geminivirus strain demarcation and nomenclature. *Arch Virol* (in press).

Garcia-Andres, S., Monci, F., Navas-Castillo, J. & Moriones, E. (2006). Begomovirus genetic diversity in the native plant reservoir *Solanum nigrum*: evidence for the presence of a new virus species of recombinant nature. *Virology* **350**, 433–442.

Garcia-Andres, S., Tomas, D. M., Sanchez-Campos, S., Navas-Castillo, J. & Moriones, E. (2007). Frequent occurrence of recombinants in mixed infections of tomato yellow leaf curl disease-associated begomoviruses. *Virology* **359**, 302–312.

Gibbs, M. J., Armstrong, J. S. & Gibbs, A. J. (2000). Sister-Scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics* **16**, 573–582.

Heath, L., van der Walt, E., Varsani, A. & Martin, D. P. (2006). Recombination patterns in aphthoviruses mirror those found in other picornaviruses. *J Virol* **80**, 11827–11832.

Inoue-Nagata, A. K., Albuquerque, L. C., Rocha, W. B. & Nagata, T. (2004). A simple method for cloning the complete begomovirus genome using the bacteriophage φ29 DNA polymerase. *J Virol Methods* **116**, 209–211.

Jain, R., Rivera, M. C. & Lake, J. A. (1999). Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci U S A* **96**, 3801–3806.

Jeske, H., Lutgemeier, M. & Preiss, W. (2001). DNA forms indicate rolling circle and recombination-dependent replication of Abutilon mosaic virus. *EMBO J* **20**, 6158–6167.

Kumar, S., Tamura, K. & Nei, M. (2004). MEGA3: integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform* **5**, 150–163.

Laufs, J., Traut, W., Heyraud, F., Matzeit, V., Rogers, S. G., Schell, J. & Gronenborn, B. (1995). In vitro cleavage and joining at the viral origin of replication by the replication initiator protein of tomato yellow leaf curl virus. *Proc Natl Acad Sci U S A* **92**, 3879–3883.

Lee, C., Grasso, C. & Sharlow, M. F. (2002). Multiple sequence alignment using partial order graphs. *Bioinformatics* **18**, 452–464.

Lefeuvre, P., Delatte, H., Naze, F., Dogley, W., Reynaud, B. & Lett, J. M. (2007). A new tomato leaf curl virus from the Seychelles archipelago. *Plant Pathol* **56**, 342.

Liu, H., Lucy, A. P., Davies, J. W. & Boulton, M. I. (2001). A single amino acid change in the coat protein of *Maize streak virus* abolishes systemic infection, but not interaction with viral DNA or movement protein. *Mol Plant Pathol* **2**, 223–228.

Mansoor, S., Briddon, R. W., Zafar, Y. & Stanley, J. (2003). Geminivirus disease complexes: an emerging threat. *Trends Plant Sci* **8**, 128–134.

Martin, D. & Rybicki, E. (2000). RDP: detection of recombination amongst aligned sequences. *Bioinformatics* **16**, 562–563.

Martin, D. P. & Rybicki, E. P. (2002). Investigation of *Maize streak virus* pathogenicity determinants using chimaeric genomes. *Virology* **300**, 180–188.

Martin, D. P., Posada, D., Crandall, K. A. & Williamson, C. (2005a). A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. *AIDS Res Hum Retroviruses* **21**, 98–102.

Martin, D. P., Williamson, C. & Posada, D. (2005b). RDP2: recombination detection and analysis from sequence alignments. *Bioinformatics* **21**, 260–262.

Martin, D. P., van der Walt, E., Posada, D. & Rybicki, E. P. (2005c). The evolutionary value of recombination is constrained by genome modularity. *PLoS Genet* **1**, e51.

Monci, F., Sanchez-Campos, S., Navas-Castillo, J. & Moriones, E. (2002). A natural recombinant between the geminiviruses *Tomato yellow leaf curl Sardinia virus* and *Tomato yellow leaf curl virus* exhibits a novel pathogenic phenotype and is becoming prevalent in Spanish populations. *Virology* **303**, 317–326.

Ndunguru, J., Legg, J. P., Aveling, T. A., Thompson, G. & Fauquet, C. M. (2005). Molecular biodiversity of cassava begomoviruses in Tanzania: evolution of cassava geminiviruses in Africa and evidence for East Africa being a center of diversity of cassava geminiviruses. *Virol J* **2**, 21.

Padidam, M., Sawyer, S. & Fauquet, C. M. (1999). Possible emergence of new geminiviruses by frequent recombination. *Virology* **265**, 218–225.

Paradis, E., Claude, J. & Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290.

Peterschmitt, M., Granier, M., Mekdoud, R., Dalmon, A., Gambin, O., Vayssières, J. F. & Reynaud, B. (1999). First report of tomato yellow leaf curl virus in Réunion Island. *Plant Dis* **83**, 303.

Posada, D. (2006). ModelTest Server: a web-based tool for the statistical selection of models of nucleotide substitution online. *Nucleic Acids Res* **34**, W700–W703.

Preiss, W. & Jeske, H. (2003). Multitasking in replication is common among geminiviruses. *J Virol* **77**, 2972–2980.

Rambaut, A. & Drummond, A. J. (2004). Tracer v1.3, Available from http://evolve.zoo.ox.ac.uk/software.html

Rojas, M. R., Gilbertson, R. L., Russel, D. R. & Maxwell, D. P. (1993). Use of degenerate primers in the polymerase chain reaction to detect whitefly-transmitted geminivirus. *Plant Dis* **77**, 340–347.

Ronquist, F. & Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574.

Rybicki, E. P. & Pietersen, G. (1999). Plant virus disease problems in the developing world. In *Advances in Virus Research*, vol. 53, pp. 127–178. Edited by K. Maramorosch, F. A. Murphy & A. J. Shatkin. San Diego, CA: Academic Press.

Saraf, M. C. & Maranas, C. D. (2003). Using a residue clash map to functionally characterize protein recombination hybrids. *Protein Eng* **16**, 1025–1034.

Saunders, K., Salim, N., Mali, V. R., Malathi, V. G., Briddon, R., Markham, P. G. & Stanley, J. (2002). Characterisation of Sri Lankan cassava mosaic virus and Indian cassava mosaic virus: evidence for acquisition of a DNA B component by a monopartite begomovirus. *Virology* **293**, 63–74.

Schnippenkoetter, W. H., Martin, D. P., Willment, J. A. & Rybicki, E. P. (2001). Forced recombination between distinct strains of *Maize streak virus*. *J Gen Virol* **82**, 3081–3090.

Seal, S. E., vandenBosch, F. & Jeger, M. J. (2006). Factors influencing begomovirus evolution and their increasing global significance: implications for sustainable control. *Crit Rev Plant Sci* **25**, 23–46.

Shimodaira, H. & Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol Biol Evol* **16**, 1114–1116.

Smith, J. M. (1992). Analyzing the mosaic structure of genes. *J Mol Evol* **34**, 126–129.

Stanley, J. (1995). Analysis of African cassava mosaic virus recombinants suggests strand nicking occurs within the conserved nonanucleotide motif during the initiation of rolling circle DNA replication. *Virology* **206**, 707–712.

Stenger, D. C., Revington, G. N., Stevenson, M. C. & Bisaro, D. M. (1991). Replicational release of geminivirus genomes from tandemly repeated copies: evidence for rolling-circle replication of a plant viral DNA. *Proc Natl Acad Sci U S A* **88**, 8029–8033.

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

Umaharan, P., Padidam, M., Phelps, R. H., Beachy, R. N. & Fauquet, C. (1998). Distribution and diversity of geminiviruses in Trinidad and Tobago. *Phytopathology* **88**, 1262–1268.

Voigt, C. A., Martinez, C., Wang, Z. G., Mayo, S. L. & Arnold, F. H. (2002). Protein building blocks preserved by recombination. *Nat Struct Biol* **9**, 553–558.

Wyatt, S. D. & Brown, J. K. (1996). Detection of subgroup III geminivirus isolates in leaf extracts by degenerate primers and polymerase chain reaction. *Phytopathology* **86**, 1288–1293.

Zhou, X., Liu, Y., Calvert, L., Munoz, C., Otim-Nape, G. W., Robinson, D. J. & Harrison, B. D. (1997). Evidence that DNA-A of a geminivirus associated with severe cassava mosaic disease in Uganda has arisen by interspecific recombination. *J Gen Virol* **78**, 2101–2111.