Journal of General Virology (2007), **88**, 3154–3165 **DOI** 10.1099/vir.0.83144-0

Genetic analysis of maize streak virus isolates from Uganda reveals widespread distribution of a recombinant variant

Betty E. Owor,[1] Darren P. Martin,[2] Dionne N. Shepherd,[1] Richard Edema,[3] Adérito L. Monjane,[1] Edward P. Rybicki,[1,4] Jennifer A. Thomson[1] and Arvind Varsani[1,4]

Correspondence
Arvind Varsani
arvind.varsani@uct.ac.za

[1]Department of Molecular and Cell Biology, University of Cape Town, Private Bag, Rondebosch 7701, South Africa

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

[3]Department of Crop Science, Faculty of Agriculture, Makerere University, PO Box 7062, Kampala, Uganda

[4]Electron Microscope Unit, University of Cape Town, Private Bag, Rondebosch 7701, South Africa

Maize streak virus (MSV) contributes significantly to the problem of extremely low African maize yields. Whilst a diverse range of MSV and MSV-like viruses are endemic in sub-Saharan Africa and neighbouring islands, only a single group of maize-adapted variants – MSV subtypes A_1–A_6 – causes severe enough disease in maize to influence yields substantially. In order to assist in designing effective strategies to control MSV in maize, a large survey covering 155 locations was conducted to assess the diversity, distribution and genetic characteristics of the Ugandan MSV-A population. PCR–restriction fragment-length polymorphism analyses of 391 virus isolates identified 49 genetic variants. Sixty-two full-genome sequences were determined, 52 of which were detectably recombinant. All but two recombinants contained predominantly MSV-A_1-like sequences. Of the ten distinct recombination events observed, seven involved inter-MSV-A subtype recombination and three involved intra-MSV-A_1 recombination. One of the intra-MSV-A_1 recombinants, designated MSV-A_1UgIII, accounted for >60 % of all MSV infections sampled throughout Uganda. Although recombination may be an important factor in the emergence of novel geminivirus variants, it is demonstrated that its characteristics in MSV are quite different from those observed in related African cassava-infecting geminivirus species.

Received 5 May 2007
Accepted 10 July 2007

INTRODUCTION

Maize streak virus (MSV; genus *Mastrevirus*, family *Geminiviridae*) is the causal agent of maize streak disease (MSD), the most significant disease of sub-Saharan Africa's most important food crop. Throughout this region, persistently low maize yields and erratic MSD epidemics dangerously undermine the health and social development of the world's poorest people. Despite the availability and commercial use of MSV-resistant maize genotypes for over 20 years, little progress has been made in controlling the virus in the maize fields of subsistence farmers where control is needed most urgently.

There is good evidence that, throughout Africa, MSD is caused by a group of viruses that all share >97 % genome-wide sequence identity with one another (Briddon *et al.*, 1994; Martin *et al.*, 2001). Although diversity between these so-called MSV-A viruses is low, there is strong phylogenetic support for their classification into at least six lineages or subtypes (named MSV-A_1–MSV-A_6; Martin *et al.*, 2001). There is also some evidence of variation in the subtype composition of MSV-A populations in different parts of Africa (Martin *et al.*, 2001).

Importantly, low continent-wide MSV diversity should vastly simplify the development of resistant maize genotypes. The situation is quite different from that experienced by biotechnologists and breeders attempting to develop

The GenBank/EMBL/DDBJ accession numbers for sequences of MSV isolates determined in this study are EF547063–EF547124 (also shown individually in Table 1).

Supplementary figures, Excel tables and alignment files are available with the online version of this paper.

cassava with resistance to the related cassava mosaic disease (CMD)-causing geminiviruses. At least seven distinct/tentative species of cassava-infecting geminivirus (CGV), each sharing <90 % genome-wide sequence identity with the others, cause CMD in sub-Saharan Africa (Fauquet *et al.*, 2003; Ndunguru *et al.*, 2005a; Bull *et al.*, 2006). Additionally, circulating inter-species recombinants of these viruses are common (Ndunguru *et al.*, 2005a; Bull *et al.*, 2006).

Although inter-species MSV recombinants have been detected, the scale of recombination, in terms of both the size of sequence tracts transferred and the genetic distances between parental viruses, appears to be much smaller than that observed in CGVs (Martin *et al.*, 2001). It is possible that the apparently striking evolutionary and demographic differences between MSV and CGVs are due to CGVs having been sampled far more thoroughly. Another explanation is that, despite many common biological features (insect transmission, largely overlapping geographical ranges and similar molecular biology), differences in the epidemiological and population genetic characteristics of the two groups are responsible for the apparently large differences in their evolutionary trajectories.

In this report, we describe the fine-scale population structure of MSV isolates sampled from Ugandan maize in 2005, and compare this with that of CGVs sampled from Kenya between 2001 and 2002. We use analysis of MSV PCR–restriction fragment-length polymorphism (PCR-RFLP) and full-genome sequence data, first to identify the major circulating MSV variants and then to determine the distribution of these in different regions of Uganda. In an attempt to identify the underlying factors responsible for vast differences in patterns of MSV and CGV recombination, we construct matched population-scale datasets and use these to infer and compare various estimates of evolutionary and population genetic parameters of MSV and two CGV species.

METHODS

Survey and collection of samples. A survey was conducted in the main maize-growing areas of Uganda, covering 23 central, north-central, eastern and western districts of the country, during May and June 2005 (in the first cropping season). In total, 460 samples, three from each field where possible, were collected in each of 155 fields. Each field was divided into two diagonals and the three samples were picked, two from the first diagonal and the third from the second diagonal. In fields with fewer than three diseased plants on the two diagonals, up to three symptomatic plants were sampled from the remainder of the field. Samples were either fresh leaves or leaf pressings made onto FTA cards as outlined previously (Owor *et al.*, 2007). The fields were approximately 10 km or more apart and contained plants ranging from 1 to 3.5 months old. Global positioning system (GPS) coordinates were captured for each sample (see Supplementary Table S1, available in JGV Online).

DNA extractions. Total DNA was extracted from 305 fresh MSV-infected leaf samples using a modified CTAB method (Kiprop *et al.*, 2002; Owor *et al.*, 2007). In addition, DNA was extracted from 155 leaf-pressed samples on FTA cards as described previously (Ndunguru *et al.*, 2005b; Owor *et al.*, 2007).

PCR-RFLP analysis. MSV DNA was detected in total plant DNA extracts by PCR using a pair of degenerate primers described previously (Willment *et al.*, 2001). Each PCR amplification product was digested with *Rsa*I, *Hpa*II, *Hae*III, *Cfo*I, *Hin*dIII, *Bam*HI and *Sau*3AI as described previously (Willment *et al.*, 2001). In addition, *in silico* RFLP analysis was performed on sequenced genomes. Derived *in vitro* and *in silico* restriction patterns were compared and categorized according to a comparative panel of all currently published MSV restriction-pattern types (Martin *et al.*, 2001; Willment *et al.*, 2001; Owor *et al.*, 2007), with newly discovered patterns being added to the comparative panel of pattern types (see Supplementary Fig. S1, available in JGV Online).

Cloning and sequencing of full-genome sequences. From the 155 fields sampled, 30 fields were chosen at random using a random-number generator and, for each chosen field, one sample with a positive PCR was selected for full-genome sequencing. An additional 32 samples were chosen for full-genome sequencing based on their unique RFLP profiles. Viral samples were amplified using phi29 DNA polymerase (TempliPhi; GE Healthcare) as described previously (Owor *et al.*, 2007). Briefly, the amplified concatemers were digested with either *Bam*HI or *Sal*I to yield approximately 2.7 kb, potentially full-length linearized viral genomes, which were gel-purified (Invisorb spin DNA extraction kit; Invitek) and cloned into pGEM-3Zf(+) (Promega). Both strands of cloned genomes were sequenced commercially (Macrogen Inc., Korea), using the primer set described previously (Owor *et al.*, 2007). GenBank accession numbers are shown in Table 1.

Sequence analyses. Sequences were assembled and edited using DNAMAN (version 5.2.9; Lynnon Biosoft) and MEGA (version 3.1; Kumar *et al.*, 2004). All available MSV and CGV genome sequences were obtained from GenBank. Sequence alignments were constructed using POA (Grasso & Lee, 2004) and edited both by eye and using the CLUSTAL_W-based (Thompson *et al.*, 1994) sub-sequence realignment tool implemented in MEGA. Phylogenetic trees were constructed by the neighbour-joining (Jules–Cantor distances, 1000 bootstrap replicates) and maximum-likelihood (Hasegawa–Kishino–Yano model, transition/transversion ratio inferred from the data and 100 bootstrap replicates) methods implemented in MEGA and PHYML (Guindon & Gascuel, 2003), respectively. Recombination was analysed using the RDP (Martin & Rybicki, 2000), GENECONV (Padidam *et al.*, 1999), BOOTSCAN (Martin *et al.*, 2005a), MAXCHI (Maynard Smith, 1992), CHIMAERA (Posada & Crandall, 2001), SISCAN (Gibbs *et al.*, 2000) and LARD (Holmes *et al.*, 1999) methods, implemented in RDP3 (Martin *et al.*, 2005c). Default settings were used throughout and only potential recombination events detected by two or more of the above methods coupled with phylogenetic evidence of recombination were considered significant. Also, the severity of Bonferroni correction during detection was minimized by only searching for recombination signals in a single sequence within groups of sequences sharing >99.7 % sequence identity. Composite likelihood estimates (CLEs) of population-scaled recombination rates and estimates of population-scaled mutation rates were inferred using the PAIRWISE component of LDHAT [finite-sites version of the Watterson θ inferred from the data, a minimum minor-allele frequency of either 0.05 (for datasets containing 28 and 68 sequences) or 0.1 (for a dataset containing 14 sequences), a grid size of 100 and a maximum ρ of 100, gene-conversion model of recombination with an average tract length of 1000 nt; McVean *et al.*, 2002]. Site-to-site variation in the CLE of the population-scaled recombination rates was assessed using the INTERVAL component of LDHAT [using pre-computed likelihoods for population-scaled mutation rate=0.01, a minimum minor-allele

Table 1. GenBank accession numbers for MSV sequences determined in this study

Isolate	GenBank accession no.
MSV-Uwak-1	EF547063
MSV-UWak-4	EF547064
MSV-UMpi-8	EF547065
MSV-UMpi-11	EF547066
MSV-UMpi-14	EF547067
MSV-UMask-18	EF547068
MSV-UMask-21	EF547069
MSV-UMask-23	EF547070
MSV-UMask-25	EF547071
MSV-UMask-26	EF547072
MSV-UMba-38	EF547073
MSV-UMba-41	EF547074
MSV-UBush-53	EF547075
MSV-UKas-63	EF547076
MSV-UKas-70	EF547077
MSV-UKas-71	EF547078
MSV-UKas-75	EF547079
MSV-UKas-76	EF547080
MSV-UKab-82	EF547081
MSV-UMub-89	EF547082
MSV-UMub-94	EF547083
MSV-ULuw-103	EF547084
MSV-ULuw-107	EF547085
MSV-ULuw-110	EF547087
MSV-UNak-111	EF547086
MSV-UNak-118	EF547088
MSV-UNak-119	EF547089
MSV-UNak-120	EF547090
MSV-UNak-123	EF547091
MSV-UNak-125	EF547092
MSV-UNak-129	EF547093
MSV-UMasin-138	EF547094
MSV-UMasin-139	EF547095
MSV-UMasin-144	EF547097
MSV-UMasin-149	EF547098
MSV-UHoi-154	EF547099
MSV-Uhoi-159	EF547100
MSV-UHoi-158	EF547101
MSV-UHoi-159	EF547102
MSV-UHoi-165	EF547103
MSV-UHoi-167	EF547104
MSV-UHoi-170	EF547105
MSV-UKib-179	EF547096
MSV-UKib-182	EF547106
MSV-UKib-188	EF547107
MSV-ULuw-192	EF547108
MSV-ULuw-196	EF547109
MSV-UMuk-203	EF547110
MSV-UJin-219	EF547111
MSV-UIga-224	EF547112
MSV-UIga-231	EF547113
MSV-UIga-235	EF547114
MSV-UIga-243	EF547116
MSV-UIga-244	EF547115
MSV-UBug-245	EF547117
MSV-UBug-248	EF547118
MSV-UBus-255	EF547119
MSV-UTor-271	EF547120
MSV-UKap-289	EF547121
MSV-UKap-292	EF547122
MSV-UMbal-304	EF547123
MSV-UMbal-308	EF547124

frequency cutoff of 0.05 or 0.01 (decided as described above), a gene-conversion model with an average tract length of 1000 nt, block penalty of 10, starting ρ of 5, 10^7 Markov chain Monte Carlo updates with sampling every 2000 updates and the first 500 samples discarded; McVean et al., 2004]. For variable recombination-rate analysis with LDHAT, we avoided analysis inaccuracies at the edges of alignments by simulating circular genome sequences. Alignments were constructed from tandem repeats of full-genome sequences, with the origins of virion-strand replication situated at positions 25 and 75 % along the alignments. Results of analyses with these alignments were then processed to exclude the first 25 % and last 25 % of point recombination-rate estimates. Departures from expectations of neutral evolution in full-genome sequences were inferred by using Tajima's D (Tajima, 1989) and Fu and Li's D* (Fu & Li, 1993) statistics, calculated by using the CONVERT component of LDHAT with no minor-allele frequency cutoff. The significance of these statistics contingent on inferred population-scaled mutation and recombination rates was tested by coalescent simulation using DnaSP (Rozas & Rozas, 1999).

RESULTS AND DISCUSSION

PCR-RFLP analysis of Ugandan MSV diversity

Mastrevirus-specific degenerate-primer PCR yielded the expected PCR amplicons of approximately 1.3 kb for 91.4 % (416) of 460 samples analysed. Of these, 387 yielded sufficient DNA (>1 μg) for RFLP analyses.

Most of the RFLPs observed have been reported previously for MSV-A isolates (Willment *et al.*, 2001; Owor *et al.*, 2007). Whilst no new patterns were found for *Hin*dIII, *Cfo*I, *Bam*HI and *Hpa*II, we observed one new pattern each for *Rsa*I and *Hae*III, and five new patterns for *Sau*3AI (see Supplementary Fig. S1, available in JGV Online). Of the 387 samples analysed, 6.20 % (24/387) had evidence of mixtures of previously described RFLP patterns. Although these samples possibly represented mixed infections, only 14 of them were included in further analyses, as it was possible to identify unambiguously the RFLP profiles of the two viruses present in these (this was achievable for these 14 samples because only one of the seven restriction-enzyme digests used per sample yielded evidence of multiple restriction patterns). In addition to *in vitro* RFLP analysis, cloning and sequencing of 62 full-length genomes (see below) allowed us to perform *in silico* RFLP analysis of these sequences. This led to the identification of seven and one previously unpublished RFLP patterns for *Hpa*II and *Hin*dIII, respectively. These patterns have subtle differences that would be difficult to distinguish on an agarose gel, but are nonetheless different variants (see Supplementary Fig. S1, available in JGV Online). Therefore,

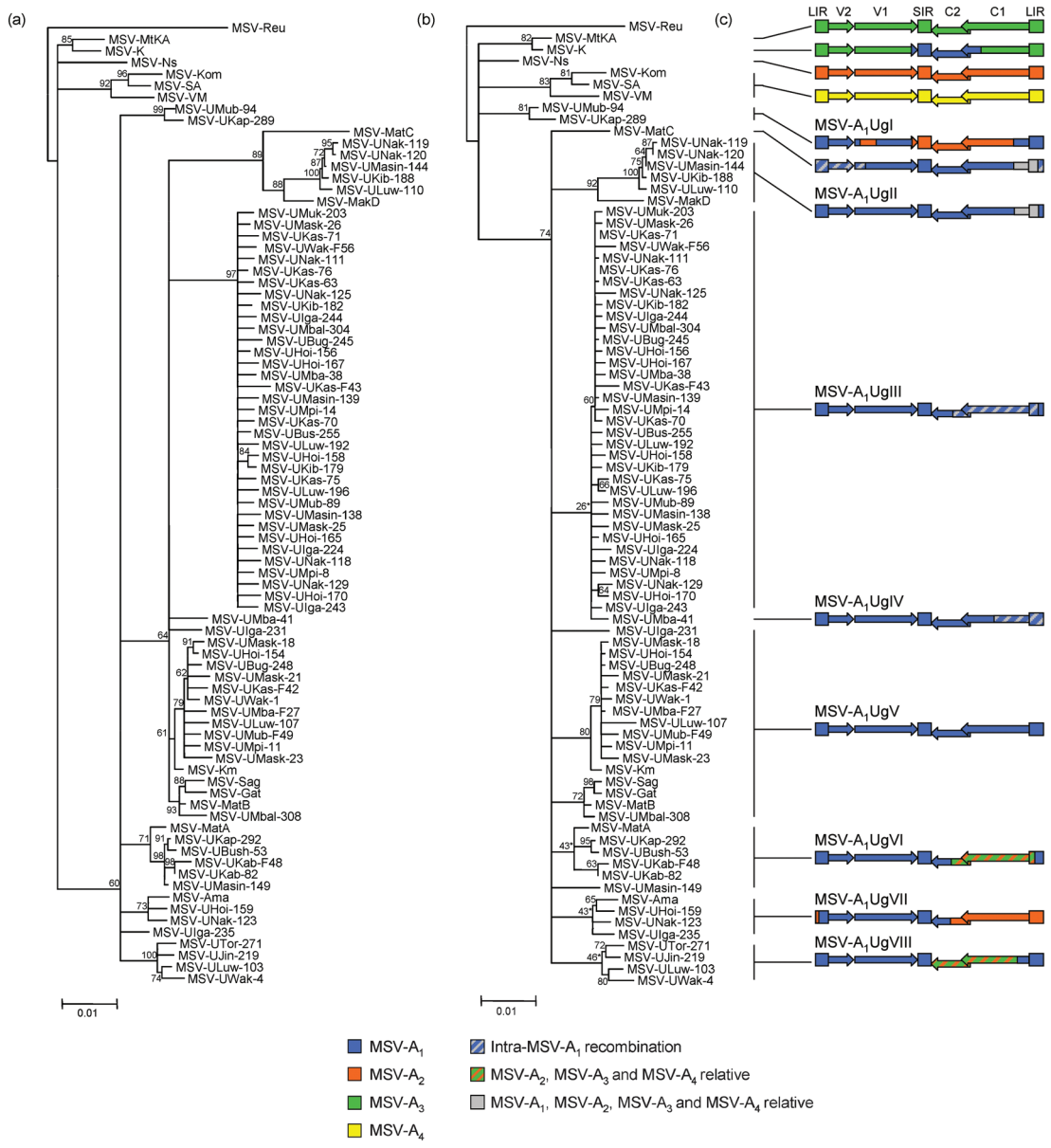

Fig. 1. Maximum-likelihood phylogenetic trees indicating possible evolutionary relationships between 84 maize-adapted maize streak virus (MSV-A) isolates. (a) Tree constructed using complete-genome sequences. The names of the Ugandan sequences contain the prefix 'MSV-U' followed by a three- or four-letter abbreviation of the sampling location name, followed by a sample number identifier that can be used to trace the sequence's GenBank accession number using the Excel spreadsheet provided as Supplementary Table S1 (available in JGV Online). All Ugandan MSV sequences other than the six with a sample number prefixed with 'F' were determined in this study. (b) Tree constructed following removal of sequence tracts that have a probable recombinant origin [indicated by the genome cartoons in (c)]. (c) Linearized genome cartoons depicting unique recombinant mosaics detected amongst the MSV-A sequences. Seven Ugandan mosaic sequences and one non-recombinant group are named MSV-A_1UgI to MSV-A_1UgVIII. Genome regions are indicated above the top cartoon: V2, movement protein gene; V1, coat protein gene; C1 and C2, replication-associated protein gene; SIR, short intergenic region; LIR, long intergenic region. Numbers associated with branches in phylogenetic trees represent the number of 100 non-parametric full maximum-likelihood bootstrap replicates supporting the existence of the branches. Other than branches indicated by an asterisk, those with $<60\,\%$ support have been collapsed. Branches marked with an asterisk in (b) have been retained wherever both (i) groupings of the same sequences above the branches have $>70\,\%$ support in (a) and (ii) where they share a common recombination mosaic. Note that certain mosaics contain evidence of recombination between MSV-A_1 and currently unsampled MSV-A subtypes. Wherever this is the case, the group of most closely related currently sampled parental subtypes is given. The colour coding indicates our tentative subtype classification of four MSV-A lineages; the MSV-A_6 subtype is uncoded and is represented here by the tree outlier MSV-Reu.

by using a combination of *in vitro* and *in silico* RFLPs, we observed new patterns for five of the seven enzymes.

In total, 49 different RFLP-pattern combinations were observed (see Supplementary Table S1, available in JGV Online). By the convention described previously (Willment *et al.*, 2001), the four most prevalent RFLP-pattern groups were AABBDAD, BCABBAA, AAABDAD and ABABBAA, respectively representing 59.34, 11.00, 3.07 and 2.56 % of the 391 isolates in Supplementary Table S1 (available in JGV Online).

Recombination and phylogenetic analyses of full-genome sequences

To analyse the range of Ugandan MSV diversity more accurately, we cloned and completely sequenced 62 MSV genomes. Amongst these were isolates representing 45 of the 49 unique RFLP-pattern combinations that we observed. These were aligned together with six previously described Ugandan MSV genome sequences (Owor *et al.*, 2007), 21 MSV genome sequences sampled elsewhere in Africa, two *Panicum* streak virus sequences, a sugarcane streak virus sequence, a sugarcane streak Reunion virus sequence and a sugarcane streak Egypt virus sequence (94 sequences in total). Preliminary phylogenetic analysis indicated that, as expected, all of the Ugandan MSV sequences were of the maize-adapted MSV-A type. More specifically, all clustered with either the MSV-A_1 or MSV-A_5 subtypes (Fig. 1a) identified previously in a 1999 survey of African MSV diversity (Martin *et al.*, 2001).

As recombination has been reported previously in MSV (Martin *et al.*, 2001), we analysed the above alignment for evidence of recombination between the 84 MSV-A viruses and the other African streak viruses (including both other MSV types and non-MSV African streak viruses). Amongst all of the MSV-A sequences, we detected only two previously reported recombination events involving MSV-A viruses as recipients of sequence from non-MSV-A sources. Both of these recombination events, described by Martin *et al.* (2001), involve exchanges of small fragments (approx. 100 bp) between MSV-A- and MSV-B-like viruses to yield MSV-VM in one case and, in the other, the ancestor of MSV-SA, MSV-Kom and MSV-VM.

Clearly, therefore, neither inter-type nor inter-species recombination is currently a major factor in Ugandan MSV-A diversification. This does not, however, discount the possibility that intra-MSV-A recombination might be an important feature of MSV-A evolution. To increase statistical power during the analysis of intra-MSV-A recombination, both the MSV-B-like portions of the MSV-Kom, -VM and -SA genomes and all of the non-MSV-A sequences were removed from the MSV alignment before rescreening for recombination.

We detected ten potential intra-MSV-A recombination events (Bonferroni-corrected P value <0.05 for at least two different recombination-detection methods coupled with phylogenetic evidence of sequence exchange), nine of which have apparently occurred in sequences ancestral to the Ugandan MSV-A sequences determined in this study (Figs 1, 2). Six of these nine events (events 1, 3, 4, 7, 8 and 10 in Fig. 2) apparently involved sequence exchanges between MSV-A_1 and viruses related most closely to those in the MSV-A_2, -A_3 and -A_4 subtype groups. Due to generally low sequence diversity amongst African MSV-A sequences and the comparatively poor sampling of all MSV-A subtypes other than MSV-A_1, it was not possible to identify the exact origins of the recombinant regions convincingly for three of the five inter-subtype recombination events (events 4, 7 and 10 in Fig. 2).

Having identified the most probable recombinant regions of all available MSV genomes, we attempted a recombination-free reconstruction of the MSV-A phylogeny. We first removed the smaller recombinant fractions of identified recombinant sequences from the alignment (i.e. replacing the tracts of recombinant sequence comprising the smallest fractions of the sampled genomes with the indel symbol '-') and then retained only those columns of the sequence alignment in which >80 % of the sequences contained sequence data. Whilst the topology of the maximum-likelihood phylogeny constructed using the resulting alignment is almost completely in agreement with that constructed using the unedited alignment, the total tree length of the former phylogeny is substantially smaller than that of the latter [compare trees in Fig. 1(a) and (b)]. This reduction in tree length is exactly what would be expected following removal of the most significant recombination signals within the dataset (Schierup & Hein, 2000).

Importantly, the sequences for MSV-MakD and MSV-MatC, formerly classified as belonging to the MSV-A_5 subtype (Martin *et al.*, 2001), cluster clearly with the MSV-A_1 subtype in both of the trees. Although both maximum-likelihood and neighbour-joining reconstructions with only the MSV-A sequences used in the study by Martin *et al.* (2001) confirmed the tree topology determined in that study (data not shown), when using our much larger MSV-A sequence dataset, both tree-construction methods indicate that the MSV-A_5 subtype is in fact a sublineage of the MSV-A_1 subtype. Although the reason for this discrepancy is unclear, MSV-MakD and MSV-MatC are both recombinants containing large tracts of MSV-A_1-like sequence [identified here and by Martin *et al.* (2001)], which would be expected to compromise the correct placement of these sequences in phylogenetic trees (Schierup & Hein, 2000; Posada & Crandall, 2002; Awadalla, 2003). Owing to the large proportion of MSV-A_1-like sequences within these two isolates and their close Ugandan relatives, we have opted to reclassify MSV-A_5 as a recombinant sublineage of MSV-A_1.

Distribution of major MSV genotypes in Uganda

Based on detected recombination patterns, the Ugandan MSV sequences were classified into eight haplotypes,

Event	Genome map	Breakpoint position		Recombinant (s)	Parental sequences		Detection method
		Begin	End		Minor	Major	
	LIR V2 V1 SIR C2 C1 LIR						
1		1588	44	MSV-Ama (MSV-A1UgVII)	MSV-Ns	MSV-UMask-18	rg**M**CSL
2		2332	590	MSV-MatC	MSV-UNak-119	MSV-UTor-271	RgM**C**S
3		2351	527	MSV-UKap-289 MSV-UMub-94	MSV-ULuw-107	MSV-Ns	rg**M**C
4		1598	2583	MSV-UBush-53 (MSV-A1UgVI)	Unsampled (MSV-SA, MSV-MtKA, MSV-Ns)	MSV-MatB	**M**cl
5		2104	2689	MSV-UMba-41	MSV-Ulga-231	MSV-UMpi-8	MC**L**
6		1616	2626	MSV-UKas-70 (MSV-A1-UgIII)	MSV-MakD	MSV-ULuw-107	r**M**c
7		2337	2626	MSV-MakD MSV-MatC (MSV-A1-UgII)	Unsampled	MSV-MatB	**m**c
8		717	1143	MSV-UKap-289 MSV-UMub-94	MSV-ULuw-107	MSV-Ns	r**S**
9		1129	1945	MSV-K	MSV-Km	MSV-MtKA	m**c**
10		1314	2400	MSV-UJin-219 (MSV-A1-UgVIII)	Unsampled (MSV-SA, MSV-MtKA, MSV-Ns)	MSV-Ulga-231	**M**C

Fig. 2. Characterization of ten recombination events detected amongst 84 maize-adapted MSV (MSV-A) full-genome sequences. Sequences bounded by recombination breakpoints are shaded on the graphical representation of MSV-A genomes. Identities of genome regions are the same as those indicated in Fig. 1. In cases where multiple recombinant genomes are indicated, they are all descendants of the same recombinant ancestor. In certain cases, groups of closely related sequences are referred to collectively by their haplotype identifiers indicated in Fig. 1. 'Minor' and 'major' parents refer to sequences related closely to those contributing the smaller and larger fractions of the recombinant's sequence, respectively. In cases where multiple parental sequences are indicated, these are all inferred to be equally close relatives of actual parental sequences. In certain cases where parental sequences are listed as 'unsampled', the sequence names given in parentheses are representatives of major MSV-A lineages that are inferred to share a more recent common ancestor with the unsampled parental virus than they do with the other parental virus. Breakpoint positions represent the boundaries of the strongest recombination signal, but are not necessarily at, or even very close to, the breakage sites that occurred during the original recombination event. In the 'detection methods' column, letters represent methods indicating the presence of recombination with >95 % (lower-case letters) and 99 % (upper-case letters) confidence: R/r, RDP; G/g, GENECONV; M/m, MAXIMUM CHI SQUARE; C/c, CHIMAERA; S/s, SISCAN; L/l, LARD. The method indicating the clearest evidence of recombination for a particular event is represented in bold.

named MSV-A_1UgI to MSV-A_1UgVIII (Fig. 1b, c), with MSV-A_1UgV representing the MSV-A_1 sequences that are not detectably recombinant. We should specify here that our haplotype-classification scheme is not intended as a serious taxonomic proposal. It is simply a convenient and evolutionarily relevant way of splitting the Ugandan virus isolates into distinguishable groups.

By using RFLP data for the remaining 321 Ugandan MSV samples [197 from this study and 124 from Owor *et al.*

(2007)] that were not analysed by full-genome sequencing, it was possible to classify each of these into one of the eight haplotype groupings.

The sampling area in Uganda was split into seven zones and the relative proportions of the different haplotypes were determined for each of these zones (Fig. 3a). Importantly, there was no significant difference in the population frequencies of different haplotypes across the seven zones ($P=0.2964$; χ^2 test with eight haplotypes × seven sampling zones), indicating that, generally, the diversity of samples collected from any one of the zones was not significantly unrepresentative of country-wide MSV diversity.

Next, we examined triplet samples collected from individual farms to determine whether there were any significant differences in MSV diversity sampled on these farms, at scales of hundreds of square metres, relative to country-wide MSV diversity, sampled on a scale of thousands of square kilometres. We observed a surprisingly high degree of intra-field diversity, with 66 % of triplicate samples from individual fields containing viruses categorized as belonging to different haplotypes. There was no significant deviation in the patterns of haplotypes observed within individual farms relative to those observed country-wide ($P=0.213$; χ^2 test with 15 three-sample haplotype combinations × seven sampling zones). This indicates that MSV diversity observed within sampling areas of a few hundred square metres is also generally not significantly unrepresentative of that observed country-wide.

Showing that the distribution of MSV diversity at these different sampling scales is not significantly different is not, however, the same thing as showing that the distribution of diversity at the different scales is significantly similar. More careful examination of the samples collected in individual farms revealed five instances where the haplotype combinations observed were reasonably improbable ($P<0.05$; 2×2 χ^2 test). Although none of these comparisons were significant following Bonferroni correction of P values (to account for the multiple tests made), we could not discount the possibility that including all three samples from each farm might introduce a sampling bias into our analysis of country-wide Ugandan MSV population structure. Therefore, only one example of each haplotype sampled in each location was considered in subsequent analyses. Whilst this method biased estimates of population representation slightly against the two most common haplotypes, it enabled more sensitive analysis of the distributions of the six rarer haplotypes. This selection process did not, however, grossly distort the overall representation of the respective haplotypes, as illustrated in Supplementary Fig. S2 (available in JGV Online).

The intra-MSV-A_1 recombinant haplotype MSV-A_1UgIII (Fig. 1) comprised 50 % or more of the MSV samples in all seven zones (Fig. 3d) and is clearly the dominant MSV variant throughout Uganda. The MSV-A_1UgV haplotype, containing all of the MSV-A_1 sequences that are not obviously recombinant, is the only other haplotype that was detected in all seven sampling zones (Fig. 3e). All of the other haplotypes were either absent or present below detectable levels in two or more of the zones. As most of these were present at close to the detection limits in the zones where they were observed, it is possible that they are all present throughout the country.

There is some indication of slight variation in MSV demography in different zones. Whilst this is particularly true for some of the rarer haplotypes, such as MSV-A_1UgVI, MSV-A_1UgVII and MSV-A_1UgVIII (Fig. 3f–h), which display >5-fold variations in population representation in different zones, there is also evidence of variation in relative population representation of the more common haplotype MSV-A_1UgV. This haplotype is at its highest prevalence in four of the five eastern zones and at its lowest prevalence in the two western zones. Statistically significant deviation from country-wide population frequencies was, however, only evident for MSV-A_1UgVI in zones 5 and 6 ($P=0.015$ and 0.042, respectively; Bonferroni-corrected 2×2 χ^2 test; Fig. 3f).

Despite the possibility of slight geographical variations in haplotype frequencies across Uganda, the fact that MSV haplotype distributions do not differ substantially over sampling scales ranging from 0.1 to 100 000 km^2 suggests strongly that MSV population structure (in Uganda at least) is highly homogeneous. This in turn implies that there are no substantial impediments to the movement of viruses throughout the country.

Comparison of the population genetic characteristics of MSV and African CGVs

To compare the MSV diversity and recombination data with those of CGVs, the only other substantially sampled African geminivirus group, we obtained all 118 African cassava-infecting begomovirus DNA-A sequences available in GenBank on 30 November 2006, aligned these using POA, edited and realigned subsections of the alignments in MEGA and identified 23 major inter-species and four intra-species recombination events using RDP3 (see the supplementary files 'Cassava.rdp' and 'Cassava.csv', available in JGV Online, for detailed results of this analysis).

Relative to the MSV dataset, the CGV sequences contain more evidence of recombination involving larger fragments of sequence between more distantly related parental viruses (Fig. 4). Unsurprisingly, these quite striking differences have been noted elsewhere (Padidam *et al.*, 1999; Martin *et al.*, 2001; Schnippenkoetter *et al.*, 2001). However, the cause(s) of these differences remain unexplored. We propose that there are three main reasons that MSV and CGVs might have such different patterns of recombination: (i) the biochemical recombination rate in MSV may be significantly lower than that found in CGV species; (ii) although mixed infections, a prerequisite for detectable recombination, have been observed in both MSV and CGVs (this study; Willment *et al.*, 2001; Vanitharani *et al.*,

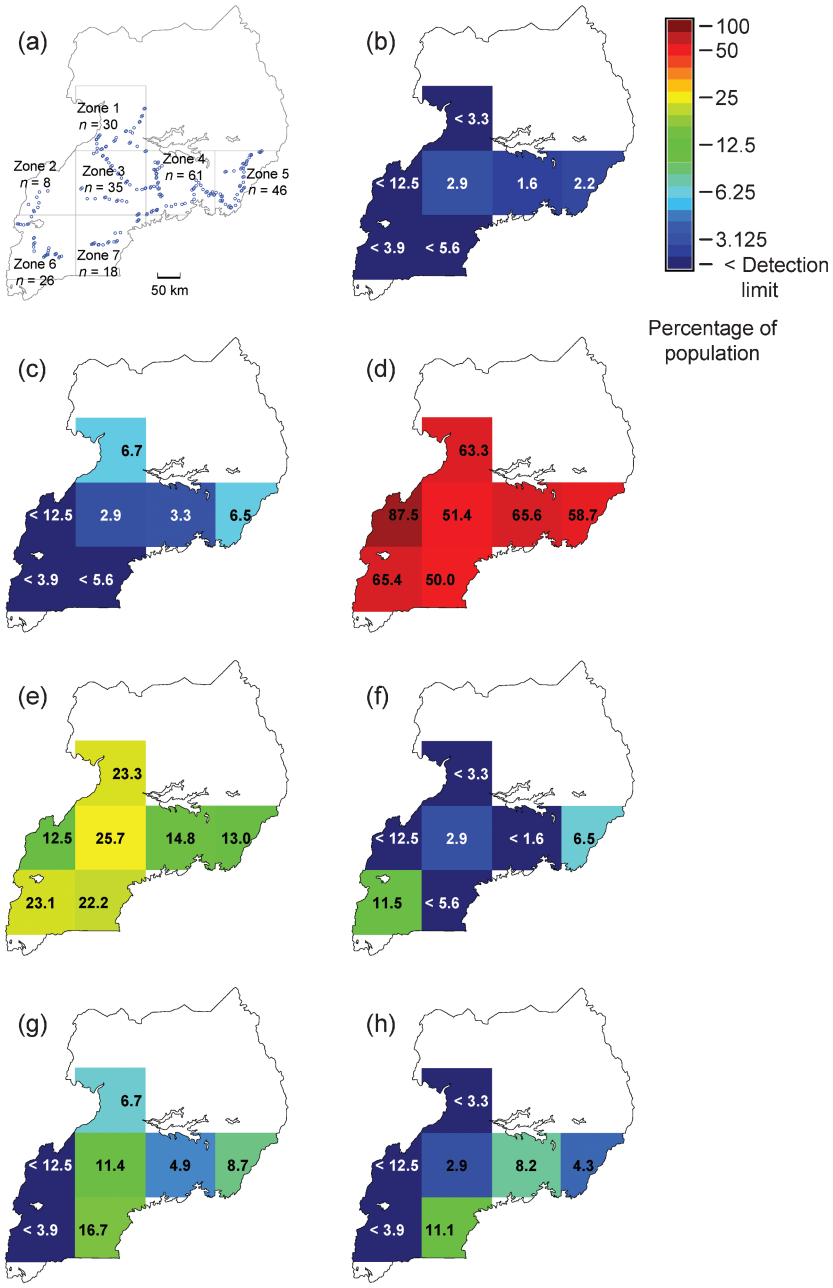

Fig. 3. Sampling locations, distributions and population representation of seven Ugandan MSV-A haplotypes. (a) Important maize-growing regions of Uganda were split into seven zones and samples were collected from 155 locations within these (blue circles). Multiple samples collected from the same location were only considered if they belonged to different haplotypes. (b–h) Distribution and population representation of: (b) MSV-A_1-UgI; (c) MSV-A_1-UgII; (d) MSV-A_1-UgIII; (e) MSV-A_1-UgV; (f) MSV-A_1-UgVI; (g) MSV-A_1-UgVII; (h) MSV-A_1-UgVIII. The colour scale is exponential (base 2) from 3.125 to 50 % and linear from 0 to 3.125 % and from 50 to 100 %.

2004), they may be much more common amongst CGVs than they are amongst MSVs; (iii) whilst the CGV species are all cassava-adapted, only one MSV strain is maize-adapted, and differences seen in the extents and prevalence of recombination in CGVs and MSV may therefore be the result of purifying selection eliminating greater proportions of MSV recombinants, particularly when these contain large tracts of sequence from viruses that are not maize-adapted (Martin & Rybicki, 2002; Martin *et al.*, 2005b).

We investigated whether there was any detectable population genetic evidence of MSV and CGVs having significantly different recombination rates. To do this, we first assembled CGV datasets with properties similar to those of our Ugandan MSV dataset. Based on the structure of the MSV dataset, we defined CGV populations arbitrarily as groups of sequences that: (i) all either shared identical inter-species recombinant mosaic structures or were not detectably inter-species recombinants; (ii) were all >96 % identical to one another; (iii) were all sampled over a geographical range similar to that of the Ugandan MSV sample; (iv) were all sampled within 2 years of one another; (v) contained more than ten completely sequenced DNA-A components. Two groups, containing 28 and 14 DNA-A sequences, respectively, were the only

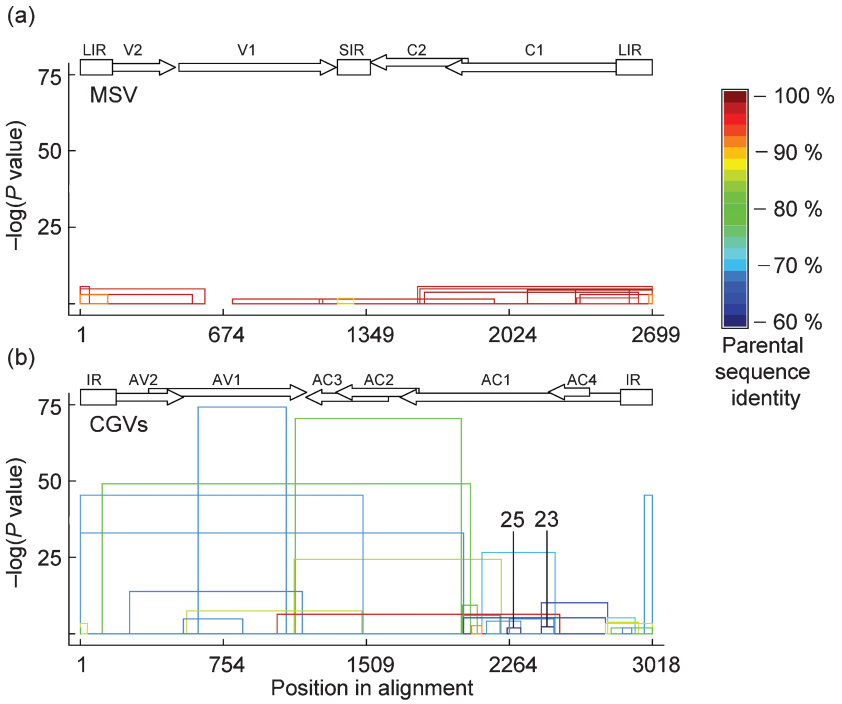

Fig. 4. Distribution and characteristics of recombination events detected amongst (a) MSV full-genome sequences and (b) DNA-A sequences of CGVs from Africa. Each rectangle represents a recombination event, with the vertical edges indicating approximate breakpoint positions and the upper horizontal edge indicating the degree of statistical support in favour of recombination. The colour of the boxes indicates the minimum degree of sequence identity shared by parental sequences at the time of the recombination event. Evidence of two recombination events involving a non-CGV parental sequence are labelled [events 23 and 25, detailed in the supporting files Cassava.csv and Cassava.rdp (available in JGV Online)]. Genome cartoons above the plots indicate the starting and ending alignment positions of virion-sense genes (arrows labelled V or AV), complementary-sense genes (arrows labelled C or AC) and intergenic regions (boxes labelled IR, LIR or SIR).

datasets that met our selection criteria. These groups contain sequences respectively classified as the East African cassava mosaic virus Uganda strain (EACMV – UG) and East African cassava mosaic Kenya virus (EACMKV). Some summary statistics indicating that these datasets are indeed very similar to the MSV dataset from a population genetic perspective are presented in Table 2.

Population-scaled recombination- and mutation-rate estimates of the different MSV and CGV populations are also presented in Table 2. It is important to make a distinction here between 'population-scaled' and 'biochemical' mutation and recombination rates. Whilst biochemical mutation and recombination rates are the rates at which nucleotide substitutions and strand-invasion events leading to recombination occur over time (such as per replication cycle or per year), the population-scaled estimates of these do not have an easily definable unit of measurement. The population-scaled mutation and recombination rates for haploid organisms evolving exclusively through genetic drift can be expressed as $2N_e u$ and $2N_e r$, respectively, where N_e is the effective population size, u is the biochemical rate at which neutral mutations occur and r is the biochemical rate at which recombination events that have no fitness effects occur. Note that whilst these population-scaled values are not actual estimates of biochemical mutation and recombination rates, both expressions contain the value $2N_e$ and therefore their ratio should be the same as the biochemical recombination- and mutation-rate ratio (the $2N_e$ part of the equations and all accessory uncertainties associated with estimating effective population sizes of viruses are cancelled out when one considers only the ratios of population-scaled mutation and recombination rates). For this relationship to hold, however, it is important that all observable mutation and recombination events considered during calculation of the

Table 2. Population genetic statistics for matched MSV and CGV datasets

Statistic	MSV	EACMV – UG	EACMKV
No. sequences	68	28	14
Nucleotide diversity (π)	0.0162	0.0124	0.0232
Population-scaled mutation rate (θ)	77.672	83.002	85.530
Population-scaled recombination rate (ρ)	4.064††	8.13	3.704†
ρ/θ	0.052	0.098	0.043
Tajima's D	−1.880††	−2.294†††	−1.102
Fu and Li's D*	−4.291†††	−3.573†††	−1.323

†$P<0.05$; ††$P<0.01$; †††$P<0.001$.

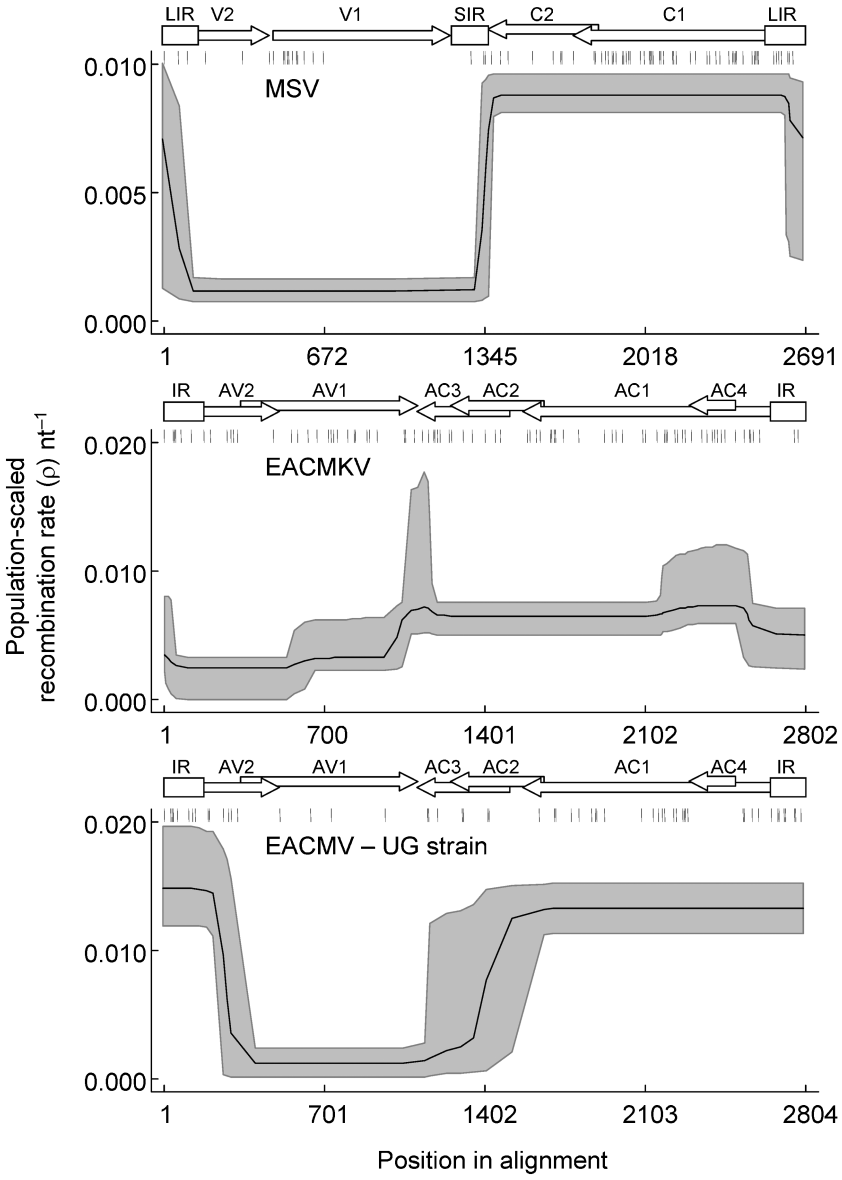

Fig. 5. Variable recombination rates along the lengths of MSV genomes and the DNA-A genome components of the CGVs East African cassava mosaic Kenya virus (EACMKV) and East African cassava mosaic virus Uganda (EACMV – UG) strain. Black lines represent mean estimates of point recombination rates determined by the reversible-jump Markov chain Monte Carlo (RJMCMC) approach implemented in the INTERVAL component of LDHAT (McVean *et al.*, 2004). Grey regions represent the 95 % credibility intervals of point recombination-rate estimates from the RJMCMC chain. Note that because of simplifying assumptions made during the approximation of likelihoods with this approach, the distribution of point recombination rates obtained from the RJMCMC chain cannot be used to infer 95 % confidence intervals of the actual point recombination rate accurately. Genome cartoons above the plots indicate the starting and ending alignment positions of virion-sense genes (arrows labelled V or AV), complementary-sense genes (arrows labelled C or AC) and intergenic regions (boxes labelled IR, LIR or SIR). Vertical lines beneath the genome cartoons indicate the locations of polymorphic sites used for the analysis.

population-scaled mutation- and recombination-rate estimates are selectively neutral. As a direct correlation between the relative neutrality of recombination events and the relatedness of parental sequences has been observed for MSV (Martin & Rybicki, 2002; Martin *et al.*, 2005b), we chose to use only very closely related groups of sequences to derive recombination- and mutation-rate estimates for both MSV and CGVs. Nevertheless, the proportions of mutation and recombination events that are neutral in these datasets are unknown and the values of the population-scaled mutation and recombination rates should be interpreted with caution.

Importantly, neither the recombination rates nor recombination-rate : mutation-rate ratios of the different datasets are substantially different from one another (Table 2). It is important that the value of this ratio for the MSV dataset falls between that determined for the two CGV datasets. Assuming that the biochemical mutation rates of MSV and CGVs are not substantially different, these data imply that the biochemical recombination rates of the different groups of viruses are also not substantially different. This indicates, therefore, that the striking differences in the types of MSV and CGV recombination events detected in nature are probably not due to large differences in the biochemical recombination rates of these viruses.

Despite their apparently similar genome-wide recombination rates, we suspected that there might be differences in regional recombination-rate variation within the MSV genomes and the CGV DNA-A components. CLEs of regional variation in the population-scaled recombination rates of all three datasets were, however, surprisingly similar (Fig. 5). In all three populations, it seems that recombina-

tion rates are significantly higher in genomic regions encoding complementary-sense genes than they are in regions encoding virion-sense genes. A potentially important clue as to the mechanistic cause of this recombination-rate imbalance may be that, within 400 nt 3′ of the virion-strand replication origin, recombination-rate estimates are at their lowest in all three datasets. Apart from indicating that similar mechanistic processes are possibly responsible for recombination-rate variation across MSV and CGV genomes, this result implies that these mechanistic processes may be features of the virion-strand replication and/or complementary-strand transcription systems.

The similarities between the MSV and CGV datasets extended to other population genetic summary statistics. For example, both Tajima's D and Fu and Li's D* statistics calculated from complete-genome sequences (and accounting for the population-scaled recombination rates calculated above) indicated statistically significant departures from neutrality in the MSV and the EACMV – UG populations (Table 2). The values of these statistics were also marginally significant for the EACMKV dataset ($P=0.09$ for Tajima's D and $P=0.06$ for Fu and Li's D*). The negative values of the D and D* statistics indicate that, given an expectation of neutrality, there is an excess of low-frequency nucleotide polymorphisms in viruses sampled from these populations. Such departures from neutrality might be caused by a range of population phenomena including, for example, sporadic cycles of population collapse and expansion, such as those characteristic of both MSV and CGV epidemiology.

It is important that both the MSV and CGV genome-sequence samples bear similar marks of population genetic processes, as this indicates that differences in the patterns of recombination events detected between the two groups are possibly not due to fundamentally different evolutionary forces acting on the viruses. It is plausible that, with respect to the evolutionary benefits of recombination, the primary differences between MSV and CGVs is that the diversity of high-fitness host-adapted genome constituents available for recombinational exchange is far greater for CGVs than it is for MSV.

Concluding remarks

Whilst our survey of Ugandan maize-infecting MSV-A isolates has revealed that the vast majority of MSD infections in the country are caused by a group of very closely related viruses, we have demonstrated that this low diversity does not necessarily equate to genetic uniformity. We have found that there is substantial evidence of genetic exchange between viruses within the MSV-A group and that a recombinant is in fact the most prevalent MSV-A variant within the country. By using recombination patterns as a means of haplotyping MSV variants, we determined that the diversity of Ugandan MSVs is remarkably constant over a wide range of sampling scales, such that viral diversity within individual farms is not significantly different from that detected across the entire country. The hypothesis that recombination is an important feature of geminivirus evolution is not new, but we have demonstrated that its characteristics are strikingly different in MSV and the related CGVs. We provide some evidence that the underlying cause of these differences is probably not that CGVs have a higher biochemical recombination rate than MSV, but rather that CGVs have recombinational access to a far greater diversity of appropriately host-adapted genome constituents. The data that we have provided will be useful in future studies involving either longitudinal monitoring of Ugandan MSV population turnover or comparative genetic analyses of large MSV population samples from different parts of the African continent. Agroinfectious constructs containing the virus genomes that we have cloned will also be useful for controlled challenges of new MSV-resistant maize genotypes currently being developed and tested for release in Uganda.

ACKNOWLEDGEMENTS

The contribution of the survey team in Uganda is gratefully acknowledged. We thank Cathal Seoghe for his constructive comments on the manuscript. This research was partially funded by the National Research Foundation (South Africa). B. E. O. is supported by the Rockefeller Foundation (USA) in partnership with the University Sciences, Humanities and Engineering Partnerships in Africa (USHEPiA); D. N. S. has a fellowship from the Claude Leon Foundation; D. P. M. is supported by the Harry Oppenheimer Trust and the Sydney Brenner Fellowship. A. V. is supported by the Carnegie Corporation of New York.

REFERENCES

Awadalla, P. (2003). The evolutionary genomics of pathogen recombination. *Nat Rev Genet* **4**, 50–60.

Briddon, R. W., Lunness, P., Chamberlin, L. C. & Markham, P. G. (1994). Analysis of the genetic variability of maize streak virus. *Virus Genes* **9**, 93–100.

Bull, S. E., Briddon, R. W., Sserubombwe, W. S., Ngugi, K., Markham, P. G. & Stanley, J. (2006). Genetic diversity and phylogeography of cassava mosaic viruses in Kenya. *J Gen Virol* **87**, 3053–3065.

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.

Fu, Y. X. & Li, W. H. (1993). Statistical tests of neutrality of mutations. *Genetics* **133**, 693–709.

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.

Grasso, C. & Lee, C. (2004). Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems. *Bioinformatics* **20**, 1546–1556.

Guindon, S. & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**, 696–704.

Holmes, E. C., Worobey, M. & Rambaut, A. (1999). Phylogenetic evidence for recombination in dengue virus. *Mol Biol Evol* **16**, 405–409.

Kiprop, E. K., Baudoin, J. P., Mwang'ombe, A. W., Kimani, P. M., Mergeai, G. & Maquet, A. (2002). Characterization of Kenyan isolates of *Fusarium udum* from pigeonpea [*Cajanus cajan* (L.) Millsp.] by cultural characteristics, aggressiveness and AFLP analysis. *J Phytopathol* **150**, 517–525.

Kumar, S., Tamura, K. & Nei, M. (2004). MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief Bioinform* **5**, 150–163.

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., Willment, J. A., Billharz, R., Velders, R., Odhiambo, B., Njuguna, J., James, D. & Rybicki, E. P. (2001). Sequence diversity and virulence in *Zea mays* of maize streak virus isolates. *Virology* **288**, 247–255.

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

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

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

McVean, G., Awadalla, P. & Fearnhead, P. (2002). A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* **160**, 1231–1241.

McVean, G. A. T., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R. & Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science* **304**, 581–584.

Ndunguru, J., Legg, J. P., Aveling, T. A., Thompson, G. & Fauquet, C. M. (2005a). 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.

Ndunguru, J., Taylor, N. J., Yadav, J., Aly, H., Legg, J. P., Aveling, T., Thompson, G. & Fauquet, C. M. (2005b). Application of FTA technology for sampling, recovery and molecular characterization of viral pathogens and virus-derived transgenes from plant tissues. *Virol J* **2**, 45.

Owor, B. E., Shepherd, D. N., Taylor, N. J., Edema, R., Monjane, A. L., Thomson, J. A., Martin, D. P. & Varsani, A. (2007). Successful application of FTA Classic Card technology and use of bacteriophage phi29 DNA polymerase for large-scale field sampling and cloning of complete maize streak virus genomes. *J Virol Methods* **140**, 100–105.

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

Posada, D. & Crandall, K. A. (2001). Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc Natl Acad Sci U S A* **98**, 13757–13762.

Posada, D. & Crandall, K. A. (2002). The effect of recombination on the accuracy of phylogeny estimation. *J Mol Evol* **54**, 396–402.

Rozas, J. & Rozas, R. (1999). DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15**, 174–175.

Schierup, M. H. & Hein, J. (2000). Consequences of recombination on traditional phylogenetic analysis. *Genetics* **156**, 879–891.

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.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595.

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.

Vanitharani, R., Chellappan, P., Pita, J. S. & Fauquet, C. M. (2004). Differential roles of AC2 and AC4 of Cassava geminiviruses in mediating synergism and suppression of posttranscriptional gene silencing. *J Virol* **78**, 9487–9498.

Willment, J. A., Martin, D. P. & Rybicki, E. P. (2001). Analysis of the diversity of African streak mastreviruses using PCR-generated RFLPs and partial sequence data. *J Virol Methods* **93**, 75–87.